

Stochastic-Process Limits

A new book published by Springer-Verlag in 2002:

- Available from [Springer](#)
- Available from [Amazon](#)
- A [brief description](#)

Copies of selected chapters from the book:

- Cover, Preface and Contents [\[Postscript\]](#) [\[PDF\]](#)
- Chapter 1: Experiencing Statistical Regularity [\[Postscript\]](#) [\[PDF\]](#)
- Chapter 2: Random Walks in Applications [\[Postscript\]](#) [\[PDF\]](#)
- Chapter 5: Heavy-Traffic Limits for Queues [\[Postscript\]](#) [\[PDF\]](#)
- Chapter 6: Unmatched Jumps in the Limit Process [\[Postscript\]](#) [\[PDF\]](#)
- Chapter 12: The Space D [\[Postscript\]](#) [\[PDF\]](#)
- Chapter 13: Useful Functions [\[Postscript\]](#) [\[PDF\]](#)

There is a 300-page [Internet Supplement](#) to the book available online.

VISIT...

LANZAROTE
Caliente.COM

Stochastic-Process Limits

An Introduction to Stochastic-Process Limits And their Application to Queues

Ward Whitt

AT&T Labs - Research
The Shannon Laboratory
Florham Park, New Jersey

Draft
June 13, 2001

Copyright ©info

Preface

0.1. What Is This Book About?

This book is about *stochastic-process limits*, i.e., limits in which a sequence of stochastic processes converges to another stochastic process. Since the converging stochastic processes are constructed from initial stochastic processes by appropriately scaling time and space, the stochastic-process limits provide a macroscopic view of uncertainty. The stochastic-process limits are interesting and important because they generate simple approximations for complicated stochastic processes and because they help explain the statistical regularity associated with a macroscopic view of uncertainty.

This book emphasizes the continuous-mapping approach to obtain new stochastic-process limits from previously established stochastic-process limits. The continuous-mapping approach is applied to obtain stochastic-process limits for *queues*, i.e., probability models of service systems or waiting lines. These limits for queues are called *heavy-traffic limits*, because they involve a sequence of models in which the offered loads are allowed to increase towards the critical value for stability. These heavy-traffic limits generate simple approximations for complicated queueing processes under normal loading and reveal the impact of variability upon queueing performance. By focusing on the application of stochastic-process limits to queues, this book also provides an introduction to heavy-traffic stochastic-process limits for queues.

0.2. In More Detail

More generally, this is a book about *probability theory* – a subject which has applications to every branch of science and engineering. Probability theory can help manage a portfolio and it can help engineer a communication

network. As it should, probability theory tells us how to compute probabilities, but probability theory also has a more majestic goal: *Probability theory aims to explain the statistical regularity associated with a macroscopic view of uncertainty.*

In probability theory, there are many important ideas. But one idea might fairly lay claim to being the central idea: That idea is conveyed by the *central limit theorem*, which explains the ubiquitous bell-shaped curve: Following the giants – De Moivre, Laplace and Gauss – we have come to realize that, under regularity conditions, a sum of random variables will be approximately normally distributed if the number of terms is sufficiently large.

In the last half century, through the work of Erdős and Kac (1946, 1947), Doob (1949), Donsker (1951, 1952), Prohorov (1956), Skorohod (1956) and others, a broader view of the central limit theorem has emerged. We have discovered that there is not only statistical regularity in the n^{th} sum as n gets large, but there also is statistical regularity in the first n sums. That statistical regularity is expressed via a stochastic-process limit, i.e., a limit in which a sequence of stochastic processes converges to another stochastic process: A sequence of continuous-time stochastic processes generated from the first n sums converges in distribution to Brownian motion as n increases. That generalization of the basic central limit theorem (CLT) is known as *Donsker's theorem*. It is also called a functional central limit theorem (FCLT), because it implies convergence in distribution for many functionals of interest, such as the maximum of the first n sums. The ordinary CLT becomes a simple consequence of Donsker's FCLT, obtained by applying a projection onto one coordinate, making the ordinary CLT look like a view from Abbott's (1952) *Flatland*.

As an extension of the CLT, Donsker's FCLT is important because it has many significant applications, beyond what we would imagine knowing the CLT alone. For example, there are many applications in Statistics: Donsker's FCLT enables us to determine asymptotically-exact approximate distributions for many test statistics. The classic example is the *Kolmogorov-Smirnov statistic*, which is used to test whether data from an unknown source can be regarded as an independent sample from a candidate distribution. The stochastic-process limit identifies a relatively simple approximate distribution for the test statistic, for any continuous candidate cumulative distribution function, that can be used when the sample size is large. Indeed, early work on the Kolmogorov-Smirnov statistic by Doob (1949) and Donsker (1952) provided a major impetus for the development of the general theory of stochastic-process limits. The evolution of that story can be

seen in the books by Billingsley (1968, 1999), Csörgő and Horváth (1993), Pollard (1984), Shorack and Wellner (1986) and van der Waart and Wellner (1996).

Donsker's FCLT also has applications in other very different directions. The application that motivated this book is the application to queues: *Donsker's FCLT can be applied to establish heavy-traffic stochastic-process limits for queues.* A heavy-traffic limit for an open queueing model (with input from outside that eventually departs) is obtained by considering a sequence of queueing models, where the input load is allowed to increase toward the critical level for stability (where the input rate equals the maximum potential output rate). In such a heavy-traffic limit, the steady-state performance descriptions, such as the steady-state queue length, typically grow without bound. Nevertheless, with appropriate scaling of both time and space, there may be a nondegenerate stochastic-process limit for the entire queue-length process, which can yield useful approximations and can provide insight into system performance. The approximations can be useful even if the actual queueing systems do not experience heavy traffic. *The stochastic-process limits strip away unessential details and reveal key features determining performance.*

We are especially interested in the scaling of time and space that occurs in these heavy-traffic stochastic-process limits. It is customary to focus attention on the limit process, which serves as the approximation, but the scaling of time and space also provides important insights. For example, the scaling may reveal a *separation of time scales*, with different phenomena occurring at different time scales. In heavy-traffic limits for queues, the separation of time scales leads to unifying ideas, such as the *heavy-traffic averaging principle* (Section 2.4.2) and the *heavy-traffic snapshot principle* (Remark 5.9.1).

We obtain these many consequences of Donsker's FCLT by applying the continuous-mapping approach: Various continuous-mapping theorems imply that convergence in distribution is preserved under appropriate functions, with the simple case being a single function that is continuous. The continuous-mapping approach is much more effective with the FCLT than the CLT because many more random quantities of interest can be represented as functions of the first n partial sums than can be represented as functions of only the n^{th} partial sum. Since many heavy-traffic stochastic-process limits for queues follow from Donsker's FCLT and the continuous-mapping approach, we see that the statistical regularity revealed by the heavy-traffic limits for queues can be regarded as a consequence of the central limit theorem.

In this book we tell the story about the expanded view of the central limit theorem in more detail. We focus on stochastic-process limits, Donsker's theorem and the continuous-mapping approach. We also put life into the general theory by providing a detailed discussion of one application — queues. We give an introductory account that should be widely accessible. To help visualize the statistical regularity associated with stochastic-process limits, we perform simulations and plot stochastic-process sample paths.

However, we hasten to point out that there already is a substantial literature on stochastic-process limits, Donsker's FCLT and the continuous-mapping approach, including two editions of the masterful book by Billingsley (1968, 1999). What distinguishes the present book from previous books on this topic is our focus on *stochastic-process limits with nonstandard scaling and nonstandard limit processes*.

An important source of motivation for establishing such stochastic-process limits for queueing stochastic processes comes from evolving communication networks: Beginning with the seminal work of Leland, Taqqu, Willinger and Wilson (1994), extensive *traffic measurements* have shown that the network traffic is remarkably bursty, exhibiting complex features such as *heavy-tailed probability distributions, strong (or long-range) dependence and self-similarity*. These features present difficult engineering challenges for network design and control; e.g., see Park and Willinger (2000) and Krishnamurthy and Rexford (2001). Accordingly, a goal in our work is to gain a better understanding of these complex features and the way they affect the performance of queueing models.

To a large extent, the complex features — the heavy-tailed probability distributions, strong dependence and self-similarity — can be *defined* through their impact on stochastic-process limits. Thus, a study of stochastic-process limits, in a sufficiently broad context, is directly a study of the complex features observed in network traffic. From that perspective, it should be clear that this book is intended as a response (but not nearly a solution) to the engineering challenge posed by the traffic measurements.

We are interested in the way complex traffic affects network performance. Since a major component of network performance is congestion (queueing effects), we abstract network performance and focus on the way the complex traffic affects the performance of queues. The heavy-traffic limits show that the complex traffic can have a dramatic impact on queueing performance! We show that there are again heavy-traffic limits with these complex features, but both the scaling and the limit process may change. As in the standard case, the stochastic-process limits reveal key features determining performance.

The heavy-tailed distributions and strong dependence can lead to stochastic-process limits with *jumps in the limit process*, i.e., stochastic-process limits in which the limit process has discontinuous sample paths. The jumps have engineering significance, because they reveal sudden big changes, when viewed in a long time scale.

Much of the more technical material in the book is devoted to establishing stochastic-process limits with jumps in the limit process, but there already are books discussing stochastic-process limits with jumps in the limit process. Indeed, Jacod and Shiryaev (1987) establish many such stochastic-process limits. To be more precise, from the technical standpoint, what distinguishes this book from previous books on this topic is our focus on stochastic-process limits with *unmatched jumps* in the limit process; i.e., stochastic process limits in which the limit process has jumps unmatched in the converging processes.

For example, we may have a sequence of stochastic processes with continuous sample paths converging to a stochastic process with discontinuous sample paths. Alternatively, before scaling, we may have stochastic processes, such as queue-length stochastic processes, that move up and down by unit steps. Then, after introducing space scaling, the discontinuities are asymptotically negligible. Nevertheless, the sequence of scaled stochastic processes can converge in distribution to a limiting stochastic process with discontinuous sample paths.

Jumps are not part of Donsker's FCLT, because Brownian motion has continuous sample paths. But the classical CLT and Donsker's FCLT do not capture all possible forms of statistical regularity that can prevail. Other forms of statistical regularity emerge when the assumptions of the classical CLT no longer hold. For example, if the random variables being summed have heavy-tailed probability distributions (which here means having infinite variance), then the classical CLT for partial sums breaks down. Nevertheless, there still may be statistical regularity, but it assumes a new form. Then there is a different FCLT in which the limit process has jumps!

But the jumps in this new FCLT are *matched jumps*; each jump corresponds to an exceptionally large summand in the sums. At first glance, it is not so obvious that unmatched jumps can arise. Thus, we might regard stochastic-process limits with unmatched jumps in the limit process as pathological, and thus not worth serious attention. Part of the interest here lies in the fact that such limits, not only can occur, but routinely do occur in interesting applications. In particular, unmatched jumps in the limit process frequently occur in heavy-traffic limits for queues in the presence of heavy-tailed probability distributions. For example, in a single-server queue,

the queue-length process usually moves up and down by unit steps. Hence, when space scaling is introduced, the jumps in the scaled queue-length process are asymptotically negligible. Nevertheless, occasional exceptionally long service times can cause a rapid buildup of customers, causing the sequence of scaled queue-length processes to converge to a limit process with discontinuous sample paths. We give several examples of stochastic-process limits with unmatched jumps in the limit process in Chapter 6.

Stochastic-process limits with unmatched jumps in the limit process present technical challenges: Stochastic-process limits are customarily established by exploiting the function space D of all right-continuous $\mathbb{R}^k$ -valued functions with left limits, endowed with the Skorohod (1956) J_1 topology (notion of convergence), which is often called “the Skorohod topology.” However, that topology does not permit stochastic-process limits with unmatched jumps in the limit process.

As a consequence, to establish stochastic-process limits with unmatched jumps in the limit process, we need to use a nonstandard topology on the underlying space D of stochastic-process sample paths. Instead of the standard J_1 topology on D , we use the M_1 topology on D , which also was introduced by Skorohod (1956). Even though the M_1 topology was introduced a long time ago, it has not received much attention. Thus, a major goal here is to provide a systematic development of the function space D with the M_1 topology and associated stochastic-process limits.

It turns out the standard J_1 topology is stronger (or finer) than the M_1 topology, so that previous stochastic-process limits established using the J_1 topology also hold with the M_1 topology. Thus, while the J_1 topology sometimes cannot be used, the M_1 topology can almost always be used. Moreover, the extra strength of the J_1 topology is rarely exploited. Thus, we would be so bold as to suggest that, *if only one topology on the function space D is to be considered, then it should be the M_1 topology.*

Even though our motivation comes from queueing models and their application to describe the performance of evolving communication networks, there are many other possible applications of stochastic-process limits with jumps in the limit process. Indeed, stochastic-process limits with jumps in the limit process can arise whenever there are abrupt changes. There are natural applications to insurance, because insurance claim distributions often have heavy tails. There also are natural applications to finance, especially in the area of risk management; e.g., related to electricity derivatives. See Embrechts, Klüppelberg and Mikosch (1997), Adler, Feldman and Taqqu (1998) and Asmussen (2000).

In some cases, the fluctuations in a stochastic process are so strong

that no stochastic-process limit is possible with a limiting stochastic process having sample paths in the function space D . In order to establish stochastic-process limits involving such dramatic fluctuations, we introduce larger function spaces than D , which we call E and F . The names are chosen to suggest a natural progression starting from the space C of continuous functions and going beyond D . We define topologies on the spaces E and F analogous to the M_2 and M_1 topologies on D . Thus we exploit our study of the M topologies on D in this later work.

Even though the special focus here is on heavy-traffic stochastic-process limits for queues allowing unmatched jumps in the limit process, many heavy-traffic stochastic-process limits for queues have no jumps in the limit process. That is the case whenever we can directly apply the continuous-mapping approach with Donsker's FCLT. Then we deduce that reflected Brownian motion can serve as an asymptotically-exact approximation for several queueing processes in a heavy-traffic limit. In the queueing chapters we show how those classic heavy-traffic limits can be established and applied. Indeed, the book is also intended to serve as a general introduction to heavy-traffic stochastic-process limits for queues.

0.3. Organization of the Book

The book has fifteen chapters, which can be roughly grouped into four parts, ordered according to increasing difficulty. The level of difficulty is far from uniform: The first part is intended to be accessible with less background. It would be helpful (necessary?) to know something about probability and queues.

The *first part*, containing the first five chapters, provides an informal introduction to stochastic-process limits and their application to queues. The first part provides a broad overview, mostly without proofs, intending to complement and supplement other books, such as Billingsley (1968, 1999).

Chapter 1 uses simulation to help the reader directly experience the statistical regularity associated with stochastic-process limits. Chapter 2 discusses applications of the random walks simulated in Chapter 1. Chapter 3 introduces the mathematical framework for stochastic-process limits. Chapter 4 provides an overview of stochastic-process limits, presenting Donsker's theorem and some of its generalizations. Chapter 5 provides an introduction to heavy-traffic stochastic-process limits for queues.

The *second part*, containing Chapters 6 – 10, shows how the unmatched jumps can arise and expands the treatment of queueing models. The first chapter, Chapter 6 uses simulation to demonstrate that there should indeed

be unmatched jumps in the limit process in several examples. Chapter 7 continues the overview of stochastic-process limits begun in Chapter 4. The remaining chapters in the second part apply the stochastic-process limits, with the continuous-mapping approach, to obtain more heavy-traffic limits for queues.

The *third part*, containing Chapters 11 – 14, is devoted to the technical foundations needed to establish stochastic-process limits with unmatched jumps in the limit process. The earlier queueing chapters draw on the third part to a large extent. The queueing chapters are presented first to provide motivation for the technical foundations.

The third part begins with Chapter 11, which provides more details on the mathematical framework for stochastic-process limits, expanding upon the brief introduction in Chapter 3. Chapter 12 focuses on the function space D of right-continuous $\mathbb{R}^k$ -valued functions with left limits, endowed with one of the nonstandard Skorohod (1956) M topologies (M_1 or M_2). As a basis for applying the continuous-mapping approach to establish new stochastic-process limits in this context, Chapter 13 shows that commonly used functions from D or $D \times D$ to D preserve convergence with the M topologies. The third part concludes with Chapter 14, which establishes heavy-traffic limits for networks of queues.

The *fourth part*, containing Chapter 15, is more exploratory. It initiates new directions for research. Chapter 15 introduces the new spaces larger than D that can be used to express stochastic-process limits for scaled stochastic processes with even greater fluctuations.

The organization of the book is described in more detail at the end of Chapter 3, in Section 3.6.

Additional material is contained in an *Internet Supplement*. The Internet Supplement has three purposes: First, it is intended to maintain a list of corrections for errors found after the book has been published. Second, it is intended to provide supporting details, such as omitted proofs, for material in the book. Third, it is intended to provide supplementary material related to the subject of the book. Pointers to the Internet Supplement will be provided throughout the book. The initial contents of the Internet Supplement appear at the end of the book in Appendix B. The Internet Supplement is available online:

<http://www.research.att.com/~wow/supplement.html>

0.4. What is Missing?

Even though this book is long, it only provides introductions to stochastic-process limits and heavy-traffic stochastic-process limits for queues.

There are several different kinds of limits that can be considered for probability distributions and stochastic processes. Here we only consider central limit theorems and natural generalizations to the functions space D . We omit other kinds of limits such as large deviation principles. For large deviation principles, the continuous-mapping approach can be applied using contraction principles. Large deviations principles can be very useful for queues; see Shwartz and Weiss (1995). For a sample of other interesting probability limits (related to the Poisson clumping heuristic), see Aldous (1989).

Even though much of the book is devoted to queues, we only discuss heavy-traffic stochastic-process limits for queues. There is a large literature on queues. Nice *general introductions to queues*, at varying mathematical levels, are contained in the books by Asmussen (1987), Cooper (1982), Hall (1991), Kleinrock (1975, 1976) and Wolff (1989).

Queueing theory is intended to aid in the *performance analysis* of complex systems, such as computer, communication and manufacturing systems. We discuss performance implications of the heavy-traffic limits, but we do not discuss performance analysis in detail. Jain (1991) and Gunther (1998) discuss the performance analysis of computer systems; Bertsekas and Gallager (1987) discuss the performance analysis of communication networks; and Buzacott and Shanthikumar (1993) and Hopp and Spearman (1996) discuss the performance analysis of manufacturing systems.

Since we are motivated by evolving communication networks, we discuss queueing models that arise in that context, but we do not discuss the context itself. For background on evolving communication networks, see Keshav (1997), Kurose and Ross (2000) and Krishnamurthy and Rexford (2001). For research on communication network performance, see Park and Willinger (2000) and recent proceedings of *IEEE INFOCOM* and *ACM SIGCOMM*:

<http://www.ieee-infocom.org/2000/>

<http://www.acm.org/pubs/contents/proceedings/series/comm/>

Even within the relatively narrow domain of *heavy-traffic stochastic-process limits for queues*, we only provide an introduction. Harrison (1985) provided a previous introduction, focusing on Brownian motion and Brownian queues, the heavy-traffic limit processes rather than the heavy-traffic limits themselves. Harrison shows how martingales and the Ito stochastic calculus can be applied to calculate quantities of interest and solve control

problems. Newell (1982) provides useful perspective as well with his focus on deterministic and diffusion approximations. Harrison and Newell show that the limit processes can be used directly as approximations without considering stochastic-process limits. In contrast, we emphasize insights that can be gained from the stochastic-process limits, e.g., from the scaling.

The subject of heavy-traffic stochastic-process limits remains a very active research topic. Most of the recent interest focuses on *networks of queues with multiple classes of customers*. A principal goal is to determine good policies for scheduling and routing. That focus places heavy-traffic stochastic-process limits in the mainstream of operations research.

Multi-class queueing networks are challenging because the obvious stability criterion – having the traffic intensity be less than one at each queue – can in fact fail to be sufficient for stability; see Bramson (1994a, b). Thus, for general multi-class queueing networks, the very definition of heavy traffic is in question. For some of the recent heavy-traffic stochastic-process limits, new methods beyond the continuous-mapping approach have been required; see Bramson (1998) and Williams (1998a,b).

Discussion of the heavy-traffic approach to multi-class queueing networks, including optimization issues, can be found in the recent books by Chen and Yao (2001) and Kushner (2001), in the collections of papers edited by Yao (1994), Kelly and Williams (1995), Kelly, Zachary and Ziedins (1996), Dai (1998), McDonald and Turner (2000) and Park and Willinger (2000), and in recent papers such as Bell and Williams (2001), Harrison (2000, 2001a,b) and Kumar (2000). Hopefully, this book will help prepare readers to appreciate that important work and extend it in new directions.

Contents

Preface	iii
0.1 What Is This Book About?	iii
0.2 In More Detail	iii
0.3 Organization of the Book	ix
0.4 What is Missing?	xi
1 Experiencing Statistical Regularity	1
1.1 A Simple Game of Chance	1
1.1.1 Plotting Random Walks	2
1.1.2 When the Game is Fair	4
1.1.3 The Final Position	10
1.1.4 Making an Interesting Game	17
1.2 Stochastic-Process Limits	21
1.2.1 A Probability Model	21
1.2.2 Classical Probability Limits	26
1.2.3 Identifying the Limit Process	29
1.2.4 Limits for the Plots	32
1.3 Invariance Principles	35
1.3.1 The Range of Brownian Motion	36
1.3.2 Relaxing the IID Conditions	39
1.3.3 Different Step Distributions	42
1.4 The Exception Makes the Rule	45
1.4.1 Explaining the Irregularity	47
1.4.2 The Centered Random Walk with $p = 3/2$	47
1.4.3 Back to the Uncentered Random Walk with $p = 1/2$	55
1.5 Summary	60
2 Random Walks in Applications	63
2.1 Stock Prices	63

2.2	The Kolmogorov-Smirnov Statistic	66
2.3	A Queueing Model for a Buffer in a Switch	70
2.3.1	Deriving the Proper Scaling	71
2.3.2	Simulation Examples	75
2.4	Engineering Significance	81
2.4.1	Buffer Sizing	81
2.4.2	Scheduling Service for Multiple Sources	86
3	The Framework for Stochastic-Process Limits	93
3.1	Introduction	93
3.2	The Space $\mathcal{P}$	94
3.3	The Space D	97
3.4	The Continuous-Mapping Approach	104
3.5	Useful Functions	106
3.6	Organization of the Book	110
4	A Panorama of Stochastic-Process Limits	115
4.1	Introduction	115
4.2	Self-Similar Processes	116
4.2.1	General CLT's and FCLT's	116
4.2.2	Self-Similarity	117
4.2.3	The Noah and Joseph Effects	120
4.3	Donsker's Theorem	122
4.3.1	The Basic Theorems	122
4.3.2	Multidimensional Versions	125
4.4	Brownian Limits with Weak Dependence	128
4.5	The Noah Effect: Heavy Tails	132
4.5.1	Stable Laws	133
4.5.2	Convergence to Stable Laws	137
4.5.3	Convergence to Stable Lévy Motion	140
4.5.4	Extreme-Value Limits	142
4.6	The Joseph Effect: Strong Dependence	144
4.6.1	Strong Positive Dependence	145
4.6.2	Additional Structure	147
4.6.3	Convergence to Fractional Brownian Motion	150
4.7	Heavy Tails Plus Dependence	157
4.7.1	Additional Structure	157
4.7.2	Convergence to Stable Lévy Motion	158
4.7.3	Linear Fractional Stable Motion	161
4.8	Summary	164

5	Heavy-Traffic Limits for Fluid Queues	167
5.1	Introduction	167
5.2	A General Fluid-Queue Model	169
5.2.1	Input and Available-Processing Processes	170
5.2.2	Infinite Capacity	171
5.2.3	Finite Capacity	174
5.3	Unstable Queues	177
5.3.1	Fluid Limits for Fluid Queues	177
5.3.2	Stochastic Refinements	181
5.4	Heavy-Traffic Limits for Stable Queues	185
5.5	Heavy-Traffic Scaling	191
5.5.1	The Impact of Scaling Upon Performance	192
5.5.2	Identifying Appropriate Scaling Functions	194
5.6	Limits as the System Size Increases	197
5.7	Brownian Approximations	201
5.7.1	The Brownian Limit	201
5.7.2	The Steady-State Distribution.	203
5.7.3	The Overflow Process	207
5.7.4	One-Sided Reflection	210
5.7.5	First-Passage Times	213
5.8	Planning Queueing Simulations	215
5.8.1	The Standard Statistical Procedure	218
5.8.2	Invoking the Brownian Approximation	219
5.9	Heavy-Traffic Limits for Other Processes	222
5.9.1	The Departure Process	222
5.9.2	The Processing Time	223
5.10	Priorities	227
5.10.1	A Heirarchical Approach	228
5.10.2	Processing Times	230
6	Unmatched Jumps in the Limit Process	233
6.1	Introduction	233
6.2	Linearly Interpolated Random Walks	235
6.2.1	Asymptotic Equivalence with M_1	236
6.2.2	Simulation Examples	237
6.3	Heavy-Tailed Renewal Processes	241
6.3.1	Inverse Processes	241
6.3.2	The Special Case with $m = 1$	244
6.4	A Queue with Heavy-Tailed Distributions	250
6.4.1	The Standard Single-Server Queue	251

6.4.2	Heavy-Traffic Limits	252
6.4.3	Simulation Examples	254
6.5	Rare Long Service Interruptions	263
6.6	Time-Dependent Arrival Rates	267
7	More Stochastic-Process Limits	273
7.1	Introduction	273
7.2	Central Limit Theorem for Processes	274
7.2.1	Hahn's Theorem	274
7.2.2	A Second Limit	279
7.3	Counting Processes	283
7.3.1	CLT Equivalence	284
7.3.2	FCLT Equivalence	285
7.4	Renewal-Reward Processes	289
8	Fluid Queues with On-Off Sources	295
8.1	Introduction	295
8.2	A Fluid Queue Fed by On-Off Sources	298
8.2.1	The On-Off Source Model	298
8.2.2	Simulation Examples	301
8.3	Heavy-Traffic Limits for the On-Off Sources	307
8.3.1	A Single Source	307
8.3.2	Multiple Sources	310
8.3.3	$M/G/\infty$ Sources	314
8.4	Brownian Approximations	316
8.4.1	The Brownian Limit	316
8.4.2	Model Simplification	319
8.5	Stable-Lévy Approximations	321
8.5.1	The RSLM Heavy-Traffic Limit	321
8.5.2	The Steady-State Distribution	325
8.5.3	Numerical Comparisons	328
8.6	Second Stochastic-Process Limits	330
8.6.1	$M/G/1/K$ Approximations	331
8.6.2	Limits for Limit Processes	337
8.7	Reflected Fractional Brownian Motion	339
8.7.1	An Increasing Number of Sources	339
8.7.2	Gaussian Input	340
8.8	Reflected Gaussian Processes	343

9	Single-Server Queues	347
9.1	Introduction	347
9.2	The Standard Single-Server Queue	349
9.3	Heavy-Traffic Limits	353
9.3.1	The Scaled Processes	353
9.3.2	Discrete-Time Processes	356
9.3.3	Continuous-Time Processes	359
9.4	Superposition Arrival Processes	364
9.5	Split Processes	368
9.6	Brownian Approximations	370
9.6.1	Variability Parameters	371
9.6.2	Models with More Structure	374
9.7	Very Heavy Tails	378
9.7.1	Heavy-Traffic Limits	379
9.7.2	First Passage to High Levels	380
9.8	An Increasing Number of Arrival Processes	383
9.8.1	Iterated and Double Limits	383
9.8.2	Separation of Time Scales	389
9.9	Approximations for Queueing Networks	393
9.9.1	Parametric-Decomposition Approximations	393
9.9.2	Approximately Characterizing Arrival Processes	398
9.9.3	A Network Calculus	399
9.9.4	Exogenous Arrival Processes	406
9.9.5	Concluding Remarks	407
10	Multi-Server Queues	411
10.1	Introduction	411
10.2	Queues with Multiple Servers	412
10.2.1	A Queue with Autonomous Service	412
10.2.2	The Standard m -Server Model	415
10.3	Infinitely Many Servers	419
10.3.1	Heavy-Traffic Limits	420
10.3.2	Gaussian Approximations	424
10.4	An Increasing Number of Servers	428
10.4.1	Infinite-Server Approximations	428
10.4.2	Heavy-Traffic Limits for Delay Models	430
10.4.3	Heavy-Traffic Limits for Loss Models	433
10.4.4	Planning Simulations of Loss Models	435

11 More on the Mathematical Framework	441
11.1 Introduction	441
11.2 Topologies	442
11.2.1 Definitions	442
11.2.2 Separability and Completeness	445
11.3 The Space $\mathcal{P}$	447
11.3.1 Probability Spaces	447
11.3.2 Characterizing Weak Convergence	448
11.3.3 Random Elements	450
11.4 Product Spaces	453
11.5 The Space D	457
11.5.1 J_2 and M_2 Metrics	457
11.5.2 The Four Skorohod Topologies	460
11.5.3 Measurability Issues	462
11.6 The Compactness Approach	464
12 The Space D	471
12.1 Introduction	471
12.2 Regularity Properties of D	472
12.3 Strong and Weak M_1 Topologies	474
12.3.1 Definitions	475
12.3.2 Metric Properties	477
12.3.3 Properties of Parametric Representations	480
12.4 Local Uniform Convergence at Continuity Points	483
12.5 Alternative Characterizations of M_1 Convergence	486
12.5.1 SM_1 Convergence	486
12.5.2 WM_1 Convergence	491
12.6 Strengthening the Mode of Convergence	493
12.7 Characterizing Convergence with Mappings	494
12.8 Topological Completeness	497
12.9 Non-Compact Domains	498
12.10 Strong and Weak M_2 Topologies	501
12.11 Alternative Characterizations of M_2 Convergence	504
12.11.1 M_2 Parametric Representations	504
12.11.2 SM_2 Convergence	505
12.11.3 WM_2 Convergence	507
12.11.4 Additional Properties of M_2 Convergence	509
12.12 Compactness	511

13 Useful Functions	515
13.1 Introduction	515
13.2 Composition	516
13.3 Composition with Centering	520
13.4 Supremum	525
13.5 One-Dimensional Reflection	529
13.6 Inverse	532
13.6.1 The Standard Topologies	533
13.6.2 The M'_1 Topology	536
13.6.3 First Passage Times	538
13.7 Inverse with Centering	540
13.8 Counting Functions	547
14 Queueing Networks	551
14.1 Introduction	551
14.2 The Multidimensional Reflection Map	555
14.2.1 A Special Case	555
14.2.2 Definition and Characterization	556
14.2.3 Continuity and Lipschitz Properties	561
14.3 The Instantaneous Reflection Map	570
14.3.1 Definition and Characterization	571
14.3.2 Implications for the Reflection Map	578
14.4 Reflections of Parametric Representations	581
14.5 M_1 Continuity Results and Counterexamples	584
14.5.1 M_1 Continuity Results	584
14.5.2 Counterexamples	587
14.6 Limits for Stochastic Fluid Networks	590
14.6.1 Model Continuity	592
14.6.2 Heavy-Traffic Limits	593
14.7 Queueing Networks with Service Interruptions	596
14.7.1 Model Definition	596
14.7.2 Heavy-Traffic Limits	600
14.8 The Two-Sided Regulator	607
14.8.1 Definition and Basic Properties	608
14.8.2 With the M_1 Topologies	612
14.9 Chapter Notes	615

15 The Spaces E and F	619
15.1 Introduction	619
15.2 Three Time Scales	620
15.3 More Complicated Oscillations	624
15.4 The Space E	629
15.5 Characterizations of M_2 Convergence in E	634
15.6 Convergence to Extremal Processes	637
15.7 The Space F	641
15.8 Queueing Applications	643
16 Bibliography	649
A Regular Variation	693
B Contents of the Internet Supplement	697

Chapter 1

Experiencing Statistical Regularity

1.1. A Simple Game of Chance

A good way to experience statistical regularity is to repeatedly play a game of chance. So let us consider a simple game of chance using a spinner. To attract attention, it helps to have interesting outcomes, such as falling into an alligator pit or winning a dash for cash (e.g., you receive the opportunity to run into a bank vault and drag out as many money bags as you can within thirty seconds). However, to focus on statistical regularity, rather than fear or greed, we consider repeated plays with a simple outcome.

In our game, the payoff in each of several repeated plays is determined by spinning the spinner. We pay a fee for each play of the game and then receive the payoff indicated by the spinner. Let the payoff on the spinner be uniformly distributed around the circle; i.e., if the angle after the spin is θ , then we receive $\theta/2\pi$ dollars. Thus our payoff on one play is U dollars, where U is a uniform random number taking values in the interval $[0, 1]$.

We have yet to specify the fee to play the game, but first let us simulate the game to see what cumulative payoffs we might receive, not counting the fees, if we play the game repeatedly. We perform the simulation using our favorite random number generator, by generating n uniform random numbers $U_1, \dots, U_n$, each taking values in the interval $[0, 1]$, and then forming

associated partial sums by setting

$$S_k \equiv U_1 + \cdots + U_k, \quad 1 \leq k \leq n,$$

and $S_0 \equiv 0$, where $\equiv$ denotes equality by definition. The n^{th} partial sum S_n is the total payoff after n plays of the game (not counting the fees to play the game). The successive partial sums form a *random walk*, with U_n being the n^{th} step and S_n being the position after n steps.

1.1.1. Plotting Random Walks

Now, using our favorite plotting routine, let us plot the random walk, i.e., the $n + 1$ partial sums S_k , $0 \leq k \leq n$, for a range of n values, e.g., for $n = 10^j$ for several values of j . This simulation experiment is very easy to perform. For example, it can be performed almost instantaneously with the statistical package *S* (or *S-Plus*), see Becker, Chambers and Wilks (1988) or Venables and Ripley (1994), using the function

```
walk <- function(j) {
  uniforms <- runif(10j)           # generate random numbers
  firstsums <- cumsum(uniforms)      # form the partial sums
  sums <- c(0, firstsums)           # include a 0th sum
  index <- order(sums) - 1          # adjust the index
  plot(index, sums) }              # do the plotting
```

Plots of the random walk with $n = 10^j$ for $j = 1, \dots, 4$ are shown in Figure 1.1. For small n , e.g., for $n = 10$, we see irregularly spaced (vertically) points increasing to the right, but as n increases, the spacing between the points becomes blurred and regularity emerges: The plots approach a straight line with slope equal to $1/2$, the mean of a single step U_k . If we look at the pictures in successive plots, ignoring the units on the axes, we see that the plots become independent of n as n increases. Looking at the plot for large n produces a macroscopic view of uncertainty.

The plotter automatically plots the random walk $\{S_k : 0 \leq k \leq n\}$ in the available space. Ignoring the units on the axes is equivalent to regarding the plot as a display in the unit square. By “unit square” we do not mean that the rectangle containing the plot is necessarily a square, but that new units can range from 0 to 1 on both axes, independent of the original units. The plotter automatically plots the random walk in the available space by

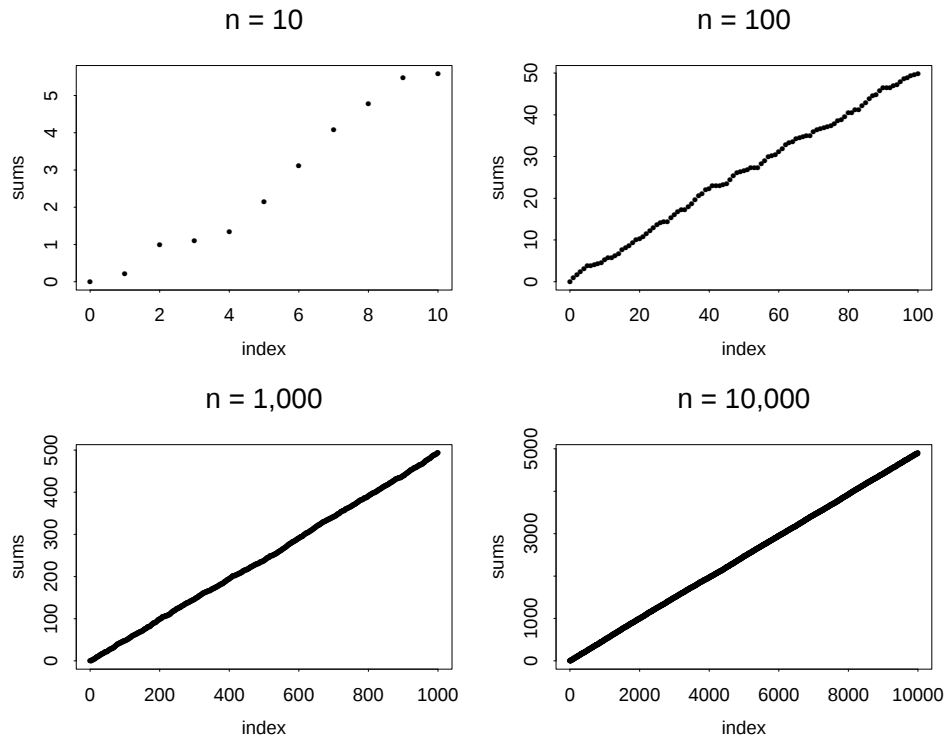


Figure 1.1: Possible realizations of the first 10^j steps of the random walk $\{S_k : k \geq 0\}$ with steps uniformly distributed in the interval $[0, 1]$ for $j = 1, \dots, 4$.

scaling time and space (the horizontal and vertical dimensions). Time is scaled by placing the $n + 1$ points $1/n$ apart horizontally. Space is scaled by subtracting the minimum and dividing by the range (assuming that the range is not zero); i.e., we interpret the plot as

$$\text{plot}(\{S_k : 0 \leq k \leq n\}) \equiv \text{plot}(\{(S_k - \min)/\text{range} : 0 \leq k \leq n\}) ,$$

where

$$\min \equiv \min(\{S_k : 0 \leq k \leq n\})$$

and

$$\text{range} \equiv \max(\{S_k : 0 \leq k \leq n\}) - \min(\{S_k : 0 \leq k \leq n\}) .$$

Combining these two forms of scaling, the plotter displays the ordered pairs $(k/n, (S_k - \min)/\text{range})$ for $0 \leq k \leq n$. With that scaling, the ordered pairs do indeed fall in the unit square. Also note that $(S_k - \min)/\text{range}$ must assume (approximately) the values 0 and 1 for at least one argument. That occurs because, without the rescaling, the plotting makes the units on the ordinate (y axis) range from the minimum value to the maximum value (approximately).

To confirm the regularity we see in Figure 1.1, we should repeat the experiment. When we repeat the experiment with different random number seeds (new uniform random numbers), the outcome for small n changes somewhat from experiment to experiment, but we always see essentially the same picture for large n . Thus the plots show regularity associated with both large n and repeated experiments.

1.1.2. When the Game is Fair

Now let us see what happens when the game is fair. Since the expected payoff is $1/2$ dollar each play of the game, the game is fair if the fee to play is $1/2$ dollar. To examine the consequences of making the game fair, we consider a minor modification of the simulation experiment above: We repeat the experiment after subtracting the mean $1/2$ from each step of the random walk; i.e., we plot the *centered random walk* (i.e., the centered partial sums $S_k - k/2$ for $0 \leq k \leq n$) for the same values of n as before.

If we consider the case $n = 10^4$, it is natural to expect to see a horizontal line instead of the line with slope $1/2$ in Figure 1.1. However, what we see is very different! Instead of a horizontal line, for $n = 10^4$ we see an irregular path, as shown in Figure 1.2.

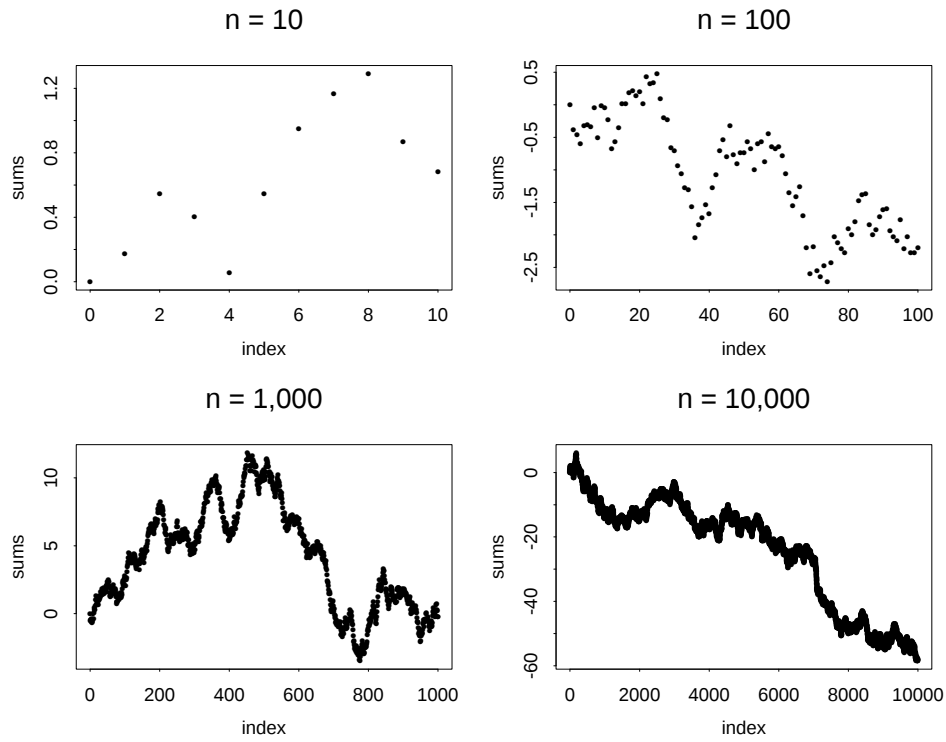


Figure 1.2: Possible realizations of the first 10^j steps of the centered random walk $\{S_k - k/2 : k \geq 0\}$ with steps uniformly distributed in the interval $[0, 1]$ for $j = 1, \dots, 4$.

We do not see the horizontal line because *the data have been automatically rescaled by the plotter*. The centering has let the plotter *blow up the picture* to show extra detail not apparent from Figure 1.1.

After centering, the range of values (the maximum minus the minimum) for the partial sums decreases dramatically. The first 10^4 uncentered partial sums assume values approximately in the interval $[0, 5000]$, whereas the first 10^4 centered partial sums all fall in the interval $[-60, 5]$. Thus, the range has decreased from 5,000 to less than 100.

At first glance, it may not be evident that there is any regularity for large n in Figure 1.2. We would hope to be able to predict what we will see if we repeat the experiment with new uniform random numbers. However, when we repeat the simulation experiment with different random number seeds, we obtain different irregular paths. To illustrate, six independent plots for $n = 10^4$ are shown in Figure 1.3. The six path samples look somewhat similar, but each is different from the others.

In Figure 1.3, just as in Figures 1.1 and 1.2, we let the plotter automatically do the scaling. Thus, the units on vertical axis change from plot to plot. We plot in this manner throughout this chapter, by design. We will show that these “automatic plots” reveal statistical regularity if we ignore the units and think of the plot as being on the unit square. But essentially the same conclusion can be drawn if we fix the units on the vertical axis. From Figure 1.3, after the fact, we can conclude that we could have fixed the units on the vertical axis, letting the values fall in the interval $[-100, 100]$. In either case, we are faced with the problem of understanding what we see.

We have arrived at a critical point, which may require us to adjust our thinking. To understand what we are seeing, we need to recognize that the irregular paths we see should be regarded as *random paths*. We then can understand that there actually is regularity underlying the six displayed paths in Figure 1.3, but it is *statistical regularity*.

We want to be able to predict what we will see when we increase n or perform additional experiments. For the uncentered random walks in Figure 1.1, we predict that the plot of $\{S_k : 0 \leq k \leq n\}$ will look like the diagonal line in the unit square for all n sufficiently large. However, for the centered random walks, the plots do not approach such a simple limit. What we should hope to predict when we repeat the experiment for the centered random walk (again ignoring the units on the axes) is the *probability distribution* of the random path. We should anticipate that the successive paths in repeated experiments will change from experiment to experiment, but we should look for a common probability distribution on the space of possible paths.

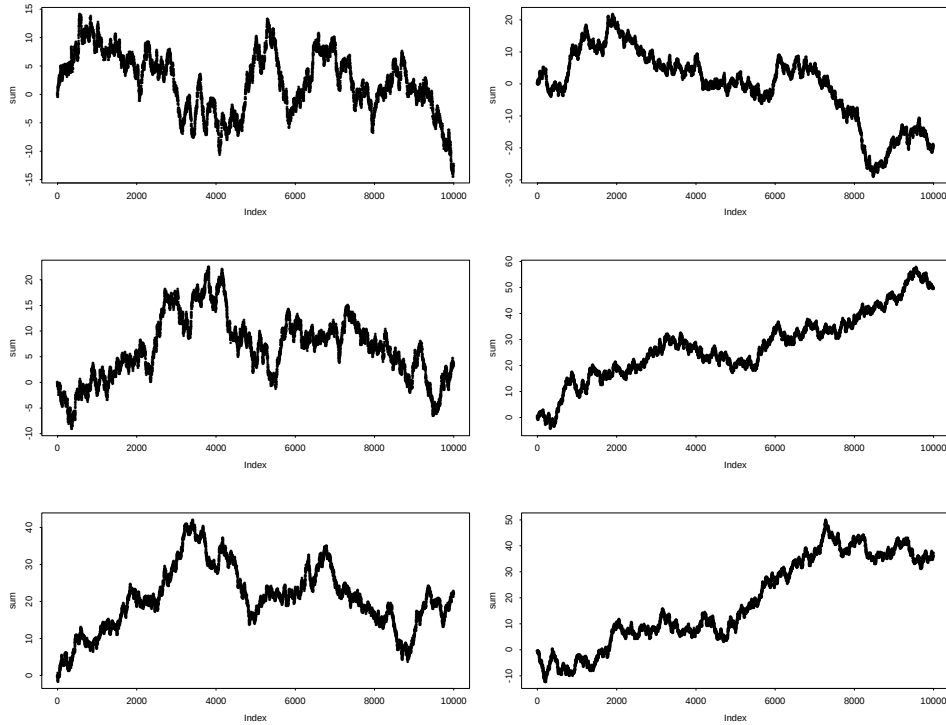


Figure 1.3: Six independent realizations of the first 10^4 steps of the centered random walk $\{S_k - k/2 : k \geq 0\}$ associated with steps uniformly distributed in the interval $[0, 1]$.

The simulation experiments suggest that, for all n sufficiently large, there tends to be a common probability distribution for the plotted random walk paths, where as before we ignore the units on the two axes or, equivalently, we regard the plot as being in the unit square. We can see part of the story when we generate new random walk paths for different values of n . For example, when we generate six centered random walk paths for $n = 10^5$ or $n = 10^6$, the plots look just like the plots in Figure 1.3. To make that clear, we plot six independent plots for the case $n = 10^6$ in Figure 1.4. As before, the units on the vertical axes change from plot to plot, but if we ignore the units on both axes, the plots in Figure 1.4 look just like the plots in Figure 1.3.

Looking at Figures 1.3 and 1.4, we should be confident about what we will see when $n = 10^8$ or $n = 10^{10}$. From Figure 1.4 and other similar plots, we see that, for n sufficiently large, the plots tend to be independent of n , provided that we ignore the units on the axes, and regard the plot as being in the unit square. Of course, as n increases, the units change on the two axes. And each new plot is a random path selected from the common probability distribution on the space of possible sample paths in the unit square.

As a consequence, we also see that the fluctuations in a smaller time scale are asymptotically negligible compared to the fluctuations in a larger time scale. Thus, for $j \geq 5$, the plots for 10^j are visually unchanged if we only keep the values at about 10^4 equally spaced indices. Indeed, such pruning of the data (reducing a data set of 10^j partial sums for $j \geq 5$ to 10^4 values) is useful to efficiently print the plots for large n .

The fact that the plots are independent of n for all n sufficiently large means that the plots tend to exhibit *self-similarity*. By self-similarity we mean that rescaled versions of the plot associated with increasing n tend to look like the original plot. More specifically, the probability distribution on the space of sample paths in the unit square tends to be unaffected by the scaling. Self-similarity will be a persistent theme; e.g., see Section 4.2.

When we consider rescaling, we can also decrease n . For instance, suppose that we consider the plot for $n = 10^7$ and select 10% of it from a subinterval of the plot. If we make a full plot of that 10% portion, then we obtain a plot for $n = 10^6$, which looks just like a random version of the original plot for $n = 10^7$. (By a “random version of the original plot” we mean that the probability distributions on the space of possible sample paths in the unit square tend to be the same.) Similarly, if we continue and select 10% of the new plot for $n = 10^6$ from any subinterval and plot it, then we obtain a plot for $n = 10^5$, which again looks like a random version of the

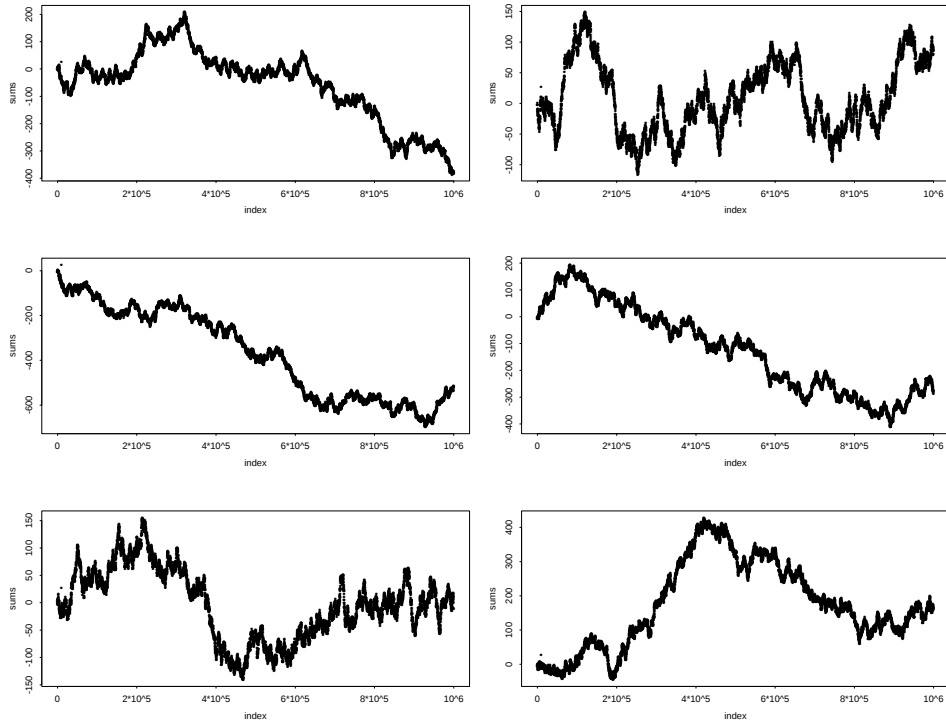


Figure 1.4: Six independent realizations of the first 10^6 steps of the centered random walk $\{S_k - k/2 : k \geq 0\}$ associated with steps uniformly distributed in the interval $[0, 1]$.

original plot for 10^7 . Of course, Figure 1.2 shows that the self-similarity for the random walks associated with decreasing n breaks down when n is too small. It is interesting to contemplate a limiting continuous-time random path that permits self-similarity without end!

1.1.3. The Final Position

It is difficult to actually see the probability distribution of the entire random path, because the path is multidimensional, but we can easily look at any one position of the random walk. For instance, suppose that we focus on the final position of the centered random walk, i.e., the single centered partial sum $S_n - n/2$ for one fixed (large) value of n .

It is evident that the final position of the centered random walk, $S_n - n/2$, changes from experiment to experiment. We find statistical regularity when we perform many independent replications of the experiment and look at the distribution of the final positions. So, let us do that.

Remark 1.1.1. *The final position and the relative final position.* For simplicity, we now want to look at the final position of the centered random walk, $S_n - n/2$, independent of the rest of the random walk. If instead we looked at the final position in the unit square, ignoring the original units, we would be looking at the *relative final position*, which must assume a value between 0 and 1. Letting $M_n \equiv \max_{1 \leq k \leq n} \{S_k - k/2\}$ and $m_n \equiv \min_{1 \leq k \leq n} \{S_k - k/2\}$, the relative final position is

$$R_n \equiv \frac{S_n - n/2 - m_n}{M_n - m_n}, \quad n \geq 1. \quad (1.1)$$

It turns out that there is statistical regularity associated with the relative final position, just as there is statistical regularity associated with the entire plot, but the relative final position is more complicated than the final position. Hence, now we focus on the final position. We discuss the relative final position in Remark 1.2.2 at the end of Section 1.2.4. ■

Suppose that we consider the final position of the centered random walk with uniform random steps for $n = 1000$, and suppose that we perform 1000 replications of the experiment. We thus obtain 1000 independent samples of the centered sum $S_{1000} - 500$. We can estimate the probability density of this distribution using the nonparametric probability density estimator *density* from *S* (with the default parameter settings). The estimated probability density of the final position $S_{1000} - 500$ is plotted in Figure 1.5.

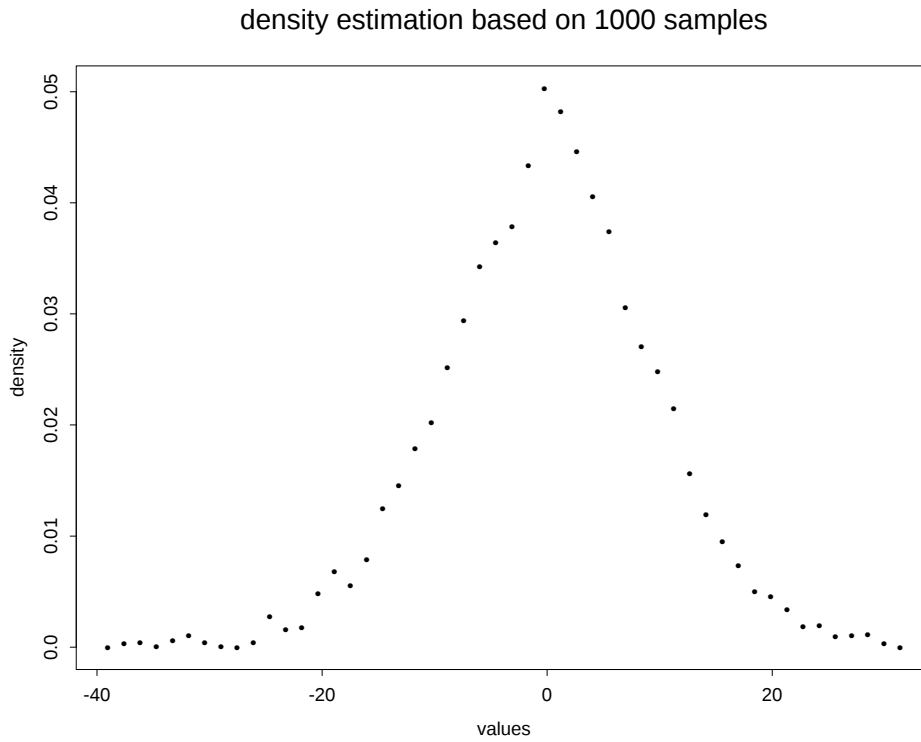


Figure 1.5: An estimate of the probability density of the final position of the random walk, obtained from 1000 independent samples of the centered partial sum $S_{1000} - 500$, where the steps U_k are uniformly distributed in the interval $[0, 1]$.

Figure 1.5 shows that nonparametric density estimation does not achieve high resolution with only a modest amount of data, but it suggests that the final position of the random walk after 1000 steps is approximately normally distributed with zero mean. That conclusion is more strongly supported by the QQ plot in Figure 1.6. The QQ plot compares the empirical distribution of the data to the normal distribution; e.g., see p. 122 of Venables and Ripley (1994). Specifically, the QQ plot compares the sorted data to the quantiles of the normal distribution. If there are n data points, then we consider the $n - 1$ normal quantiles z_k , where

$$P(N(0, 1) \leq z_k) = k/n, \quad 1 \leq k \leq n - 1,$$

with $N(m, \sigma^2)$ denoting a random variable with a normal (or Gaussian) distribution having mean m and variance σ^2 . When $n = 1,000$, the normal quantiles range from -3.1 to $+3.1$, with there being more quantiles near 0 than at the extremes. (Since we focus on the shape of the QQ plot, the QQ plot compares the distributions independent of location and scale; e.g., the shape of the QQ plot is independent of the mean and variance of the reference normal distribution.)

The near-linear plot in Figure 1.6 is approximately the same as the QQ plot for 1000 independent samples from a normal distribution. To make that clear, a QQ plot of a sample of 1000 observations from a normal distribution (with the same mean and variance) is also shown in Figure 1.6. Again the units are different in the two plots, because the range of values differs from sample to sample. The linearity that holds except for the tails strongly indicates that the final positions are indeed normally distributed.

But, in order to fairly draw that conclusion, we need more experience with QQ plots. We become more confident of the conclusion when we repeat these experiments a number of times; then we can observe the statistical variability in the QQ plots. We also gain confidence when we make QQ plots of various non-normal distributions; then we can see how departures from normality are reflected in the plots. When you think hard about the figures, they become invitations to perform additional experiments. Our main point here is that analysis with the QQ plots indicates that the final position of the centered random walk is indeed approximately normally distributed.

That conclusion is also supported by density estimates based on more data. To illustrate how the density estimates perform as a function of sample size, we display the estimates of the probability density of the same final position $S_{1000} - 500$ based on 10^j samples for $j = 2, \dots, 5$ in Figure 1.7 (again using the nonparametric density estimator *density* from *S* with the default

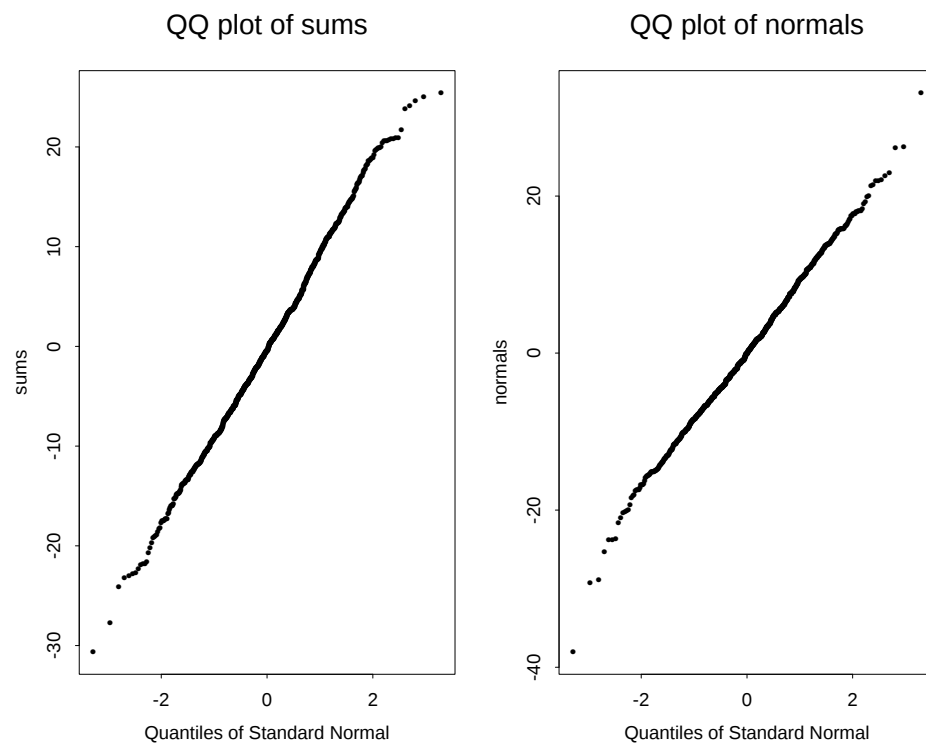


Figure 1.6: Two QQ plots of 1000 samples: the first for the sums, i.e., the final positions $S_{1000} - 500$ of the centered random walk, and the second for a normal distribution.

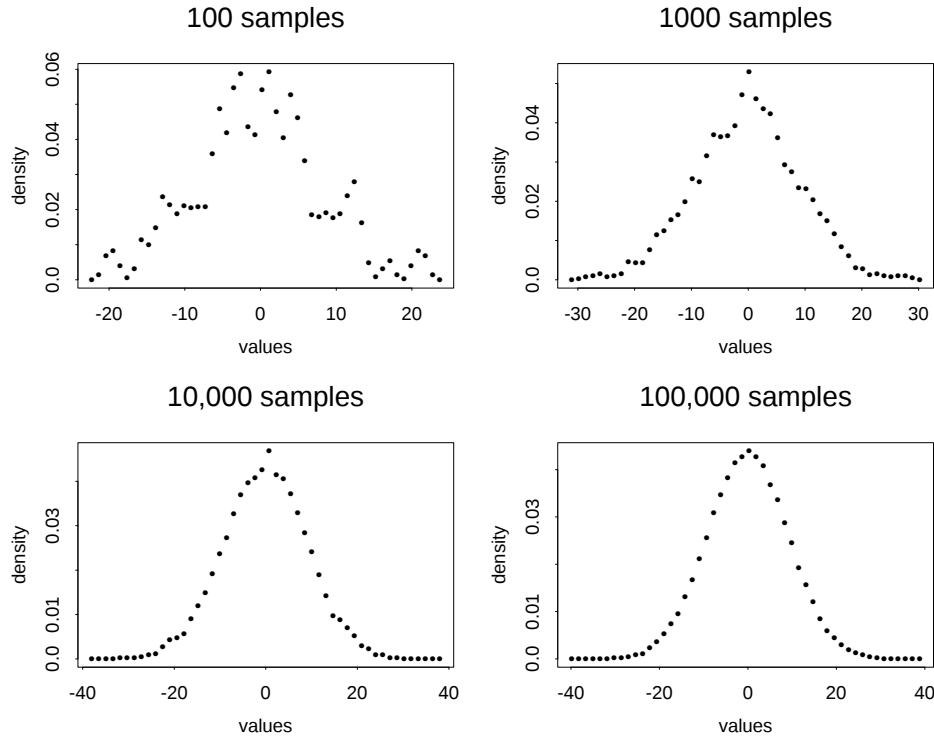


Figure 1.7: Estimates of the probability density of the final position of the random walk, obtained from 10^j independent samples of the centered partial sum $S_{1000} - 500$ for $j = 2, \dots, 5$, for the case in which the steps U_k are uniformly distributed in the interval $[0, 1]$, based on the nonparametric density estimator *density* from S .

parameter settings). Essentially the same plots are obtained for independent samples from normal distributions. From Figure 1.7, it is evident that the density estimates converge to a normal pdf as $n \rightarrow \infty$. For more on density estimation, see Devroye (1987).

It is not our purpose to delve deeply into statistical issues, but it is worth remarking that we obtain new interesting plots, like the random walk plots, when we do. Our brief examination of the distribution of the final position of the random walk suggests looking for a more precise statistical test to determine whether or not the final position of the random walk is indeed approximately normally distributed. To evaluate whether some data can be regarded as an independent sample any specified probability distribution, it

is natural to carefully investigate how the empirical distribution of a sample from that probability distribution tends to differ from the underlying probability distribution itself.

Recall that the *cumulative distribution function* (cdf) F of a random variable X is the function

$$F(t) \equiv P(X \leq t) \quad \text{for } t \in \mathbb{R}.$$

Similarly, the *empirical cdf* of a data set of size n is the proportion $F_n(t)$ of the n data points that are less than or equal to t , as a function of t .

The idea, then, is to look at the *difference* between a cdf and the empirical cdf obtained from an independent sample from that cdf. Moreover, it is natural to consider how that difference behaves as the sample size increases. Once we have made such a study, we can use the established behavior of samples from the specified probability distribution to *test* whether or not data from an unknown source can reasonably be regarded as a sample from the candidate probability distribution.

Example 1.1.1. *The empirical cdf of uniform random numbers.* To illustrate, we now consider the difference between the empirical cdf associated with n uniform random numbers on the interval $[0, 1]$ and the uniform cdf itself. Since the uniform cdf is $F(t) = t, 0 \leq t \leq 1$, we now want to plot $F_n(t) - t$ versus t for $0 \leq t \leq 1$. Since the function $F_n(t) - t, 0 \leq t \leq 1$, is a function of a continuous variable, the plotting is less routine than for the random walk. However, the empirical cdf F_n has special structure, making it possible to do the plotting quite easily. In particular, to do the plotting, let $U_k^{(n)}, 1 \leq k \leq n$, be the *order statistics* associated with the uniform random numbers $U_1, \dots, U_n$, i.e., $U_k^{(n)}$ is the k^{th} smallest of the uniform random numbers. Note that

$$F_n(U_k^{(n)}) = k/n \quad \text{and} \quad F_n(U_k^{(n)}-) = (k-1)/n,$$

$F_n(0) = 0$ and $F_n(1) = 1$, where $F_n(t-)$ is the left limit of the function F_n at t . Thus we can plot $F_n(t) - t$ versus t by plotting the points $(0, 0)$, $(1, 0)$, $(U_k^{(n)}, (k-1)/n - U_k^{(n)})$ and $(U_k^{(n)}, k/n - U_k^{(n)})$, $1 \leq k \leq n$, and connecting the points by lines (i.e., performing linear interpolation).

Plots for $n = 10^j$ for $j = 1, \dots, 4$ are shown in Figure 1.8. The plots in Figure 1.8 look much like the plots of the uncentered random walks, but there is a subtle difference that can be confirmed by further replications of the experiment. Unlike before, here the final position is 0 just like the initial

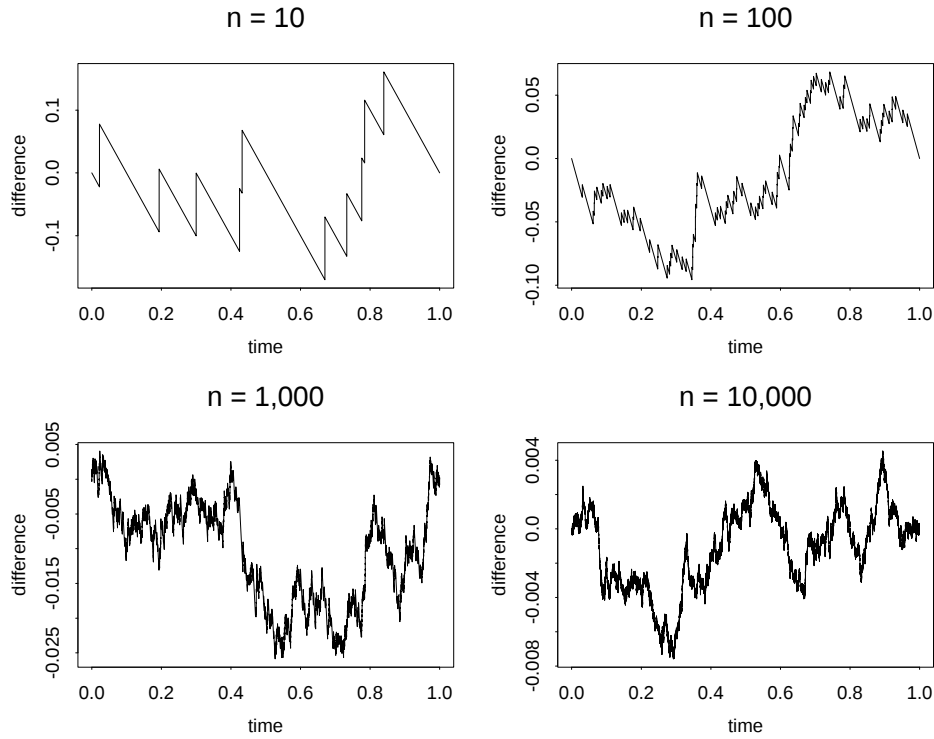


Figure 1.8: The difference between the empirical cdf and the actual cdf for samples of size 10^j from the uniform distribution over the interval $[0, 1]$ for $j = 1, \dots, 4$.

position. That makes sense as well, because both the empirical cdf and the actual cdf must assume the common value 1 at the right endpoint.

It turns out that there is statistical regularity in the empirical cdf's just like there is in the random walks. As before, the plots look the same for all sufficiently large n . Moreover, except for having the final position be 0, the plots look just like the random-walk plots. More generally, this example illustrates that statistical analysis is an important source of motivation for stochastic-process limits. We discuss this example further in Section 2.2. There we show how to develop a statistical test applicable to any continuous cdf, including the normal cdf that is of interest for the final position of the random walk.

1.1.4. Making an Interesting Game

We have digressed from our original game of chance to consider the statistical regularity observed in the plots, which of course really is our main interest. But now let us return for a moment to the game of chance.

A gambling house cannot afford to make the game fair. The gambling house needs to charge a fee greater than the expected payoff in order to make a profit. What would be a good fee for the gambling house to charge?

From the perspective of the gambling house, one might think the larger the fee the better, but the players presumably have the choice of whether or not to play. If the gambling house charges too much, few players will want to play. The fee should be large enough for the gambling house to make money, but small enough so that potential players will want to play. We take that to mean that the individual players should have a good chance of winning.

One might think that those objectives are inconsistent, but they are not. The key to achieving those objectives is the realization that *the player and the gambling house experience the game in different time scales*. An individual player might contemplate playing the game 100 times on a single day, while the gambling house might offer the game to hundreds or thousands of players on each of many consecutive days.

Thus, the player might evaluate his experience by the possible outcomes from about 100 plays of the game, while the gambling house might evaluate its experience by the possible outcomes from something like $10^4 - 10^6$ plays of the game. What we need, then, is a fee close enough to \$0.50 that the player has a good chance of winning in 100 plays, while the gambling house receives a good reliable return over $10^4 - 10^6$ games.

A reasonable fee might be \$0.51, giving the gambling house a 1 cent or 2% advantage on each play. (Gambling houses actually tend to take more, which shows the appeal of gambling despite the odds.) To see how the \$0.51 fee works, let us consider the possible experiences of the player and the gambling house. In Figure 1.9 we plot six independent realizations of a player's position during 100 plays of the game when there is a fee of \$0.51 for each play. The game looks pretty interesting for the player from Figure 1.9. The player has a reasonable chance of winning. Indeed, the player wins in plots 3 and 5, and finishes about even in plot 2. How do things look for the gambling house?

To see how the gambling house fares, we should look at the net payoffs over a much larger number of games. Hence, in Figures 1.10 and 1.11 we plot

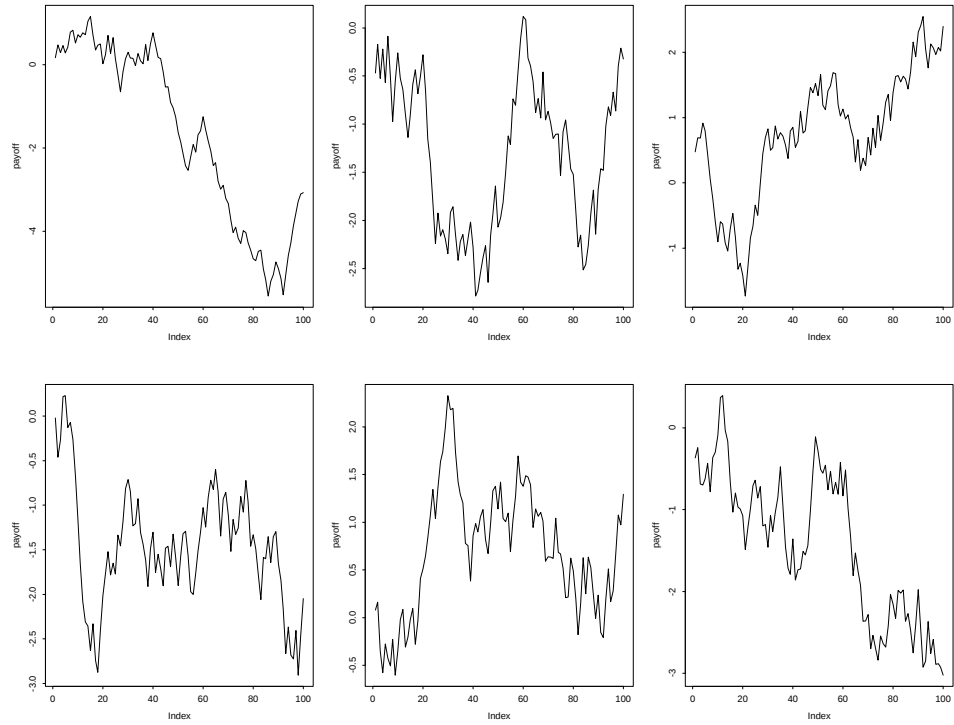


Figure 1.9: Six possible realizations of the first 100 net payoffs, positions of the random walk $\{S_k - 0.51k : k \geq 0\}$, with steps U_k uniformly distributed in the interval $[0, 1]$ and a fee of \$0.51.

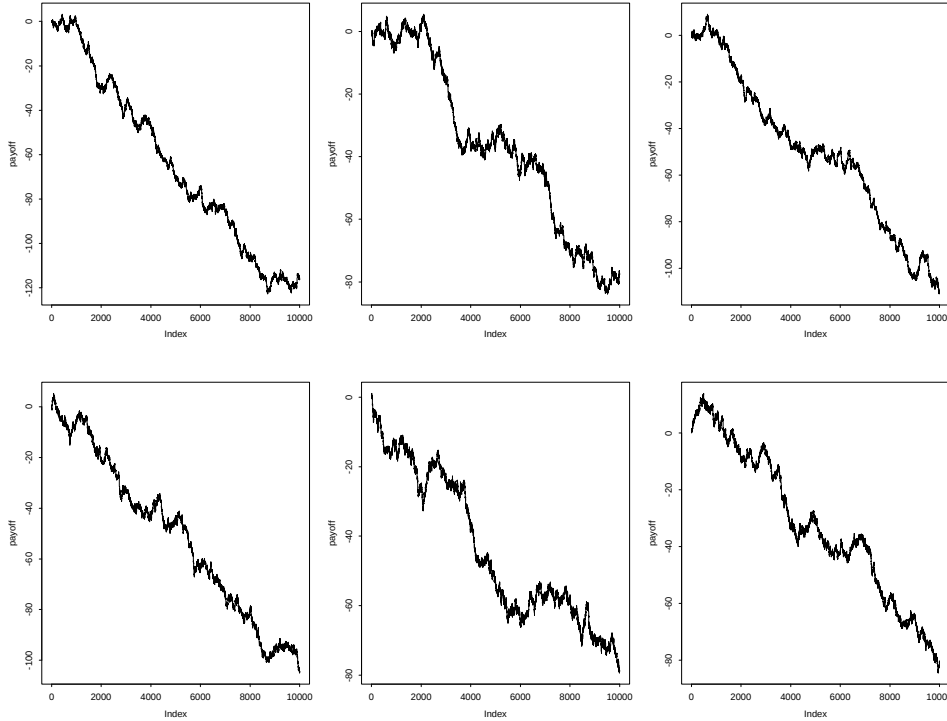


Figure 1.10: Possible realizations of the first 10^4 net payoffs (steps of the random walk $\{S_k - 0.51k : k \geq 0\}$ with steps U_k uniformly distributed in the interval $[0, 1]$).

six independent realizations of a player's position during 10^4 and 10^6 plays of the game. As before, we let the plotter automatically do the scaling, so that the units on the vertical axes change from plot to plot. But that does not alter the conclusions. In these larger time scales, we see that the player consistently loses money, so that a profit for the gambling house becomes essentially a sure thing. When we increase the number of plays to 10^6 , there is little randomness left. That is shown in Figure 1.11. Further repetitions of the experiment confirm these observations. We again see the regularity associated with a macroscopic view of uncertainty.

Above we picked a candidate fee out of the air. We could instead be more systematic. For example, we might seek the largest fee such that the player satisfies some criteria indicating a good experience. Letting the fee for each game be f , we might want to constrain the probability p that a

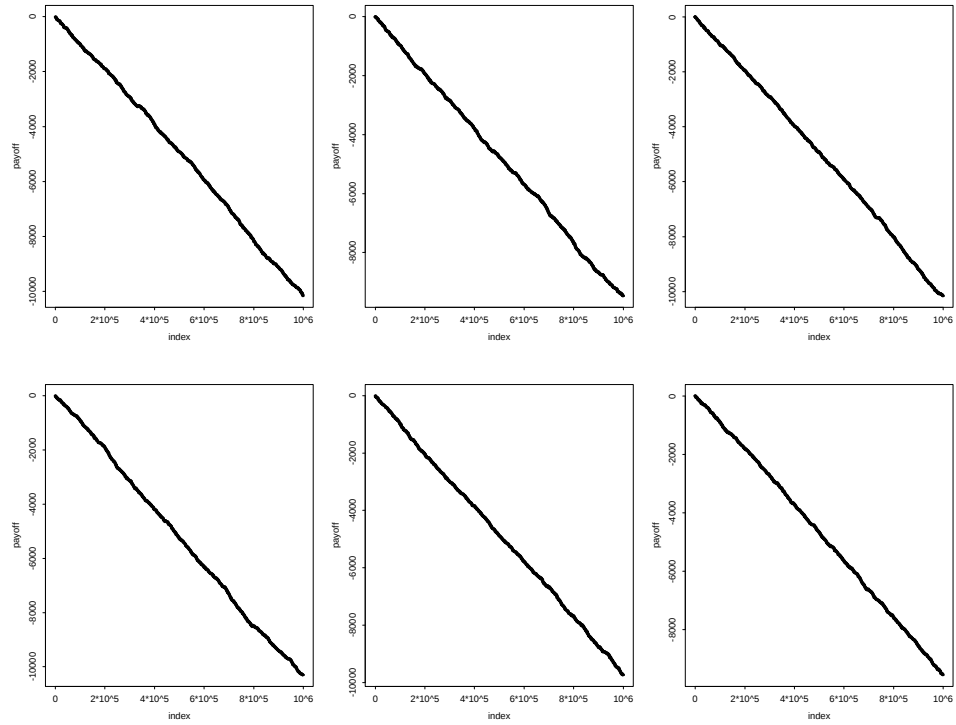


Figure 1.11: Possible realizations of the first 10^6 net payoffs (steps of the random walk $\{S_k - 0.51k : k \geq 0\}$ with steps U_k uniformly distributed in the interval $[0, 1]$).

player wins at least a certain amount w , i.e., by requiring that

$$P(S_{100} - f(100) \geq w) \geq p .$$

Given such a formulation, we can determine the optimal fee f , i.e., the maximum fee f such that the constraint is satisfied, which is attained when the probability just equals p .

As noted at the outset, when we consider making the game interesting, we might well conclude that a uniform payoff distribution for each play is boring. We might want to have the possibility of much larger positive and/or negative payoffs on one play. It is easy to devise more interesting games with different payoff distributions, but the statistical regularity associated with large numbers observed above tends to be the same. Readers are invited to make their own games and look at the net payoffs for 10^j plays for various values of j .

An extreme case that is often attractive is to have, like a lottery, some small chance of a very large payoff. However, with independent trials, as determined by successive spins of the spinner, the gambling house faces the danger of having to make too many large payoffs. Such large losses are avoided in lotteries by not letting the game be based on independent trials. In a lottery only a few prizes are awarded (and possibly shared) so that the people running the lottery are guaranteed a positive return. However, an insurance company cannot control the outcomes so tightly, so that careful analysis of the possible outcomes is necessary; e.g., see Embrechts, Klüppelberg and Mikosch (1997). We too will be interested in the possibility of exceptionally large values in random events.

1.2. Stochastic-Process Limits

The plots we have looked at indicate that there is statistical regularity associated with large n , i.e., with large sample sizes. We now want to understand *why* we see what we see, and what we will see in other related situations. For that purpose, we turn to probability theory; see Ross (1993) and Feller (1968) for introductions.

1.2.1. A Probability Model

We can use probability theory to explain what we have seen in the random walk plots. The first step is to introduce an appropriate mathematical model: Assuming that our random number generator is working properly

(an important issue, which we will not address, e.g., see p. 123 of Venables and Ripley (1994), L'Ecuyer(1998a,b) and references cited there), the observed values U_k , $1 \leq k \leq n$, should be distributed approximately as the first n values from a sequence of *independent and identically distributed* (IID) random variables uniformly distributed on $[0, 1]$ (defined on an underlying probability space). Indeed, the model fit is usually so good that there is a tendency to identify the mathematical model with the physical experiment (a mistake), but since the model fit is so good, we need not doubt that the mathematical conclusions are applicable.

Remark 1.2.1. *Mathematics and the physical world.* It is important to realize that a physical phenomenon, a mathematical model of that physical phenomenon and a simulation of that mathematical model are three different things. But, if the mathematical model is well chosen, the three may be closely related. In particular, a mathematical model, whether simulated or analyzed, may provide useful descriptions of the physical phenomenon.

We are interested in mathematical queueing models because of their ability to explain queueing phenomena, but we should not expect a perfect match. For example, mathematical models often succeed by exploiting the infinite, even though the physical phenomenon is finite. Random numbers generated on a computer are inherently finite, and yet simulations based on random numbers can be well described by mathematical models exploiting the infinite.

Here, we perform stochastic simulations to reveal statistical regularity, and we introduce and analyze mathematical models to explain that statistical regularity. We expect to capture key features, but we do not expect a perfect fit. We want the the mathematics to explain key features observed in the simulations, and we want the simulations to confirm key features predicted by the mathematics. ■

With that attitude, let us consider the probability model consisting of a sequence of IID uniform random numbers. Within the context of that probability model, we want to formulate stochastic-process limits suggested by the plots. First, we see that as n increases the plotted random walk ceases to look discrete. For all sufficiently large n , the plotted random walk looks like a function of a continuous variable. Thus it is natural to seek a continuous-time representation of the original discrete-time random walk. We can do that by considering the associated continuous-time process $\{S_{\lfloor t \rfloor} : t \geq 0\}$, where $\lfloor \cdot \rfloor$ is the *floor function*, i.e., $\lfloor t \rfloor$ denotes the greatest integer less than or equal to t . If we also want to introduce centering,

then we do the centering first, and instead consider the centered process $\{S_{[t]} - m_{[t]} : t \geq 0\}$ for appropriate centering constant m , which here is $1/2$. Thus the continuous-time representation of the random walk is a step function, which coincides with the random walk at integer arguments.

However, the step function is not the only possible continuous-time representation of the random walk. We could instead form a process with continuous sample paths by connecting the points by lines, i.e., by performing a *linear interpolation*. Then, instead of $S_{[t]}$, we consider

$$\tilde{S}(t) \equiv (t - [t])S_{[t]+1} + (1 + [t] - t)S_{[t]} \quad \text{for all } t \geq 0, \quad (2.1)$$

and similarly if we do centering. (With centering, we do the centering before doing the linear interpolation.) Possible initial segments of the two continuous-time processes associated with the discrete-time (uncentered) random walk for the case $n = 10$ are shown in Figure 1.12. (The vertical lines in the plot are not really part of the step function.) Even though the 10 random walk steps are the same for both continuous-time representations, the two initial segments of the continuous-time stochastic processes look very different in Figure 1.12. However, for large n , plots of the two continuous-time representations of the discrete-time random walk look virtually identical. To make that important point clear, we plot the two continuous-time representations of the same discrete-time centered random walk (same sample paths) for $n = 10^j$ for $j = 1, \dots, 4$ in Figure 1.13. Figure 1.13 shows that the two alternative representations indeed look the same for all n sufficiently large. Thus, when we focus on the random-walk plots for large n , we regard the two alternatives as equivalent. For our remaining discussion here, though, we will only discuss the step functions.

We now want to scale time and space (the horizontal and vertical dimensions in the plots). Note that the plotter scales time by putting the $n + 1$ random walk values in a region of fixed width. Thus, if we let 1 be the available width of the plot, then the $n + 1$ random walk values are spaced $1/n$ apart. Equivalently, time is scaled automatically by the plotting routine by multiplying time t by n , i.e., by replacing t with nt . Then, for each n , we only look at the process for t in the closed interval $[0, 1]$. The final position of the random walk for any n corresponds to $t = 1$.

We can also consider the space scaling in the same way. We can let 1 be the available height of the plot. Then the plotter automatically scales space by subtracting the minimum value and dividing by the range of the plotted values. Unfortunately, however, the range is random. Moreover, there is a complicated dependence between the path and its range. In formulating a

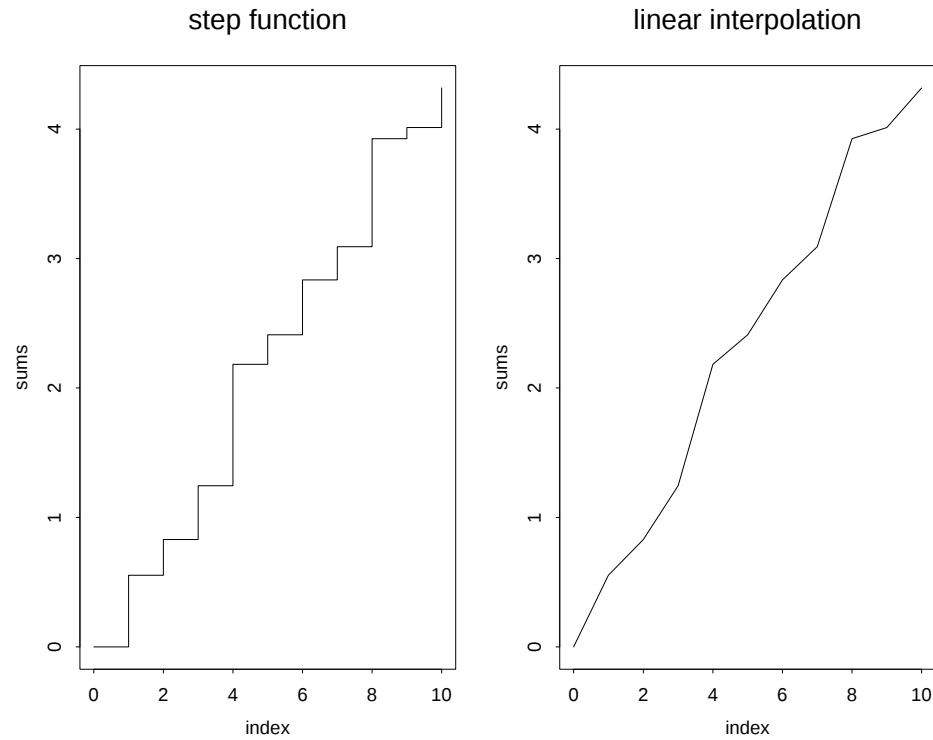


Figure 1.12: Possible initial segments of the two continuous-time stochastic processes constructed from one realization of an uncentered random walk with uniform steps for the case $n = 10$. The step-function representation appears on the left, while the linear-interpolation representation appears on the right.

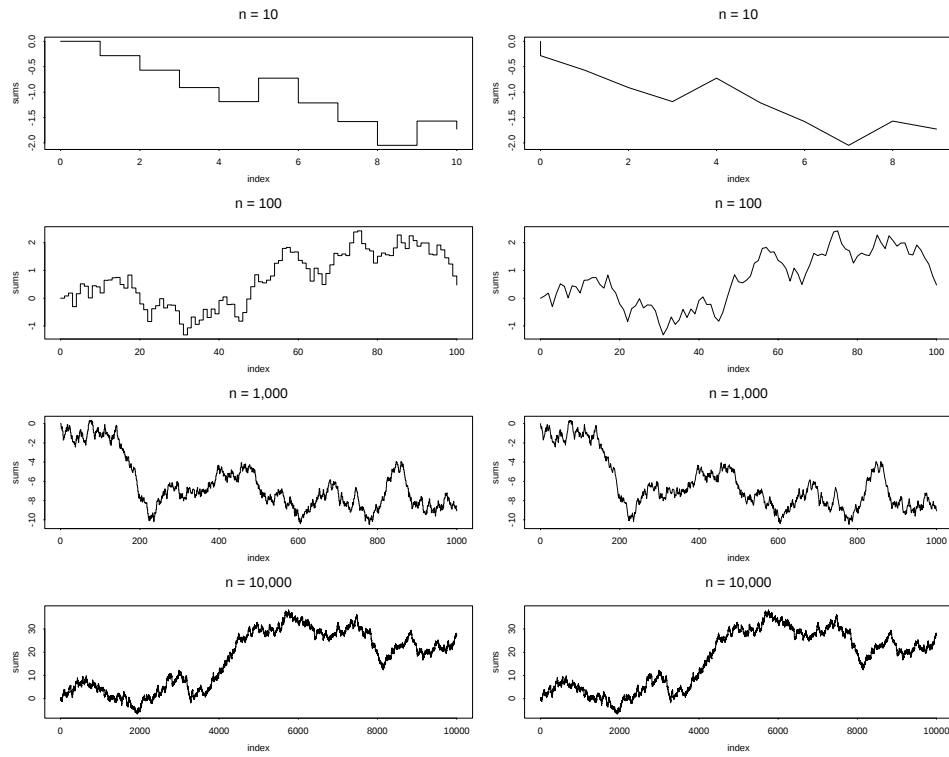


Figure 1.13: The two alternative continuous-time representations of a discrete-time process constructed from common realizations of a centered random walk with uniform steps for $n = 10^j$ with $j = 1, \dots, 4$. The step-function representation appears on the left, while the linear-interpolation representation appears on the right.

stochastic-process limit, it is natural to try to perform the space scaling, like the time scaling, with a deterministic function of n . With such deterministic space scaling, we hope to achieve a nondegenerate limit as $n \rightarrow \infty$, but one for which the range is allowed to remain random. In the limit as $n \rightarrow \infty$, we will achieve essentially the same thing as the plots if the normalized range converges to a nondegenerate random limit.

What we do, then, is scale space by dividing by c_n , where $\{c_n : n \geq 1\}$ is a sequence of (deterministic) real numbers with $c_n \rightarrow \infty$ as $n \rightarrow \infty$. That is, for each n , we form the stochastic process

$$\mathbf{S}_n(t) \equiv c_n^{-1}(S_{[nt]} - m([nt])), \quad 0 \leq t \leq 1. \quad (2.2)$$

We then want to find an appropriate sequence $\{c_n : n \geq 1\}$ so that

$$\{\mathbf{S}_n(t) : 0 \leq t \leq 1\} \rightarrow \{\mathbf{S}(t) : 0 \leq t \leq 1\} \quad \text{as } n \rightarrow \infty, \quad (2.3)$$

where $\mathbf{S} \equiv \{\mathbf{S}(t) : 0 \leq t \leq 1\}$ is an appropriate limit process with t ranging over the interval $[0, 1]$ and $\rightarrow$ in (2.3) is an appropriate mode of convergence. When we have a limit as in (2.3), we have a *stochastic-process limit*.

1.2.2. Classical Probability Limits

Classical probability limits help explain the statistical regularity we have seen. First, referring to the asymptotically linear plots in Figures 1.1 and 1.11, the *strong law of large numbers* (SLLN) implies that the scaled partial sums $n^{-1}S_n$ approach the mean m as $n \rightarrow \infty$ with probability 1 (w.p.1); e.g., see Chapter X of Feller (1968), Chapter VII of Feller (1971) and Chapter 5 of Chung (1974). (In Figure 1.1 the mean is $1/2$; in Figure 1.11 the mean is -0.01 .)

As an easy consequence of the SLLN, we can also conclude that

$$n^{-1}S_{[nt]} \rightarrow mt \quad \text{w.p.1} \quad \text{as } n \rightarrow \infty$$

for each $t > 0$. Moreover, the pointwise convergence can actually be extended to uniform convergence over bounded intervals:

$$\{n^{-1}S_{[nt]} : 0 \leq t \leq 1\} \rightarrow \{mt : 0 \leq t \leq 1\} \quad \text{w.p.1} \quad \text{as } n \rightarrow \infty,$$

uniformly in t for t in the interval $[0, 1]$. In other words,

$$\sup\{|n^{-1}S_{[nt]} - mt| : 0 \leq t \leq 1\} \rightarrow 0 \quad \text{w.p.1} \quad \text{as } n \rightarrow \infty.$$

Thus, in the setting of Figure 1.1, the limit (2.3) holds without centering (with $m = 0$) for $c_n = n$ with the limit process $\mathbf{S}$ being the line with slope $1/2$ (the one-step mean) defined over the interval $[0, 1]$. In this case, the mode of convergence in (2.3) is convergence w.p.1 on a space of functions with the uniform distance

$$\|x_1 - x_2\| \equiv \sup \{|x_1(t) - x_2(t)| : 0 \leq t \leq 1\} .$$

In this case, the stochastic-process limit is called a *functional strong law of large numbers* (FSLN). Interestingly, the SLLN and the FSLN are actually equivalent; see Theorem 3.2.1 in the Internet Supplement.

Next, turning to the plots of the centered random walks, with centering by the mean, in Figures 1.2, 1.3 and 1.4, we can appeal to the *central limit theorem* (CLT). The CLT implies that

$$(\sigma^2 n)^{-1/2}(S_n - nm) \Rightarrow N(0, 1) \quad \text{as } n \rightarrow \infty , \quad (2.4)$$

where $m \equiv EU_k = 1/2$ is the mean and $\sigma^2 \equiv Var U_k = 1/12$ is the variance of the uniform summand U_k , $\Rightarrow$ denotes convergence in distribution and the *standard* normal random variable $N(0, 1)$ has cdf

$$\Phi(t) \equiv P(N(0, 1) \leq x) \equiv \int_{-\infty}^x (2\pi)^{-1/2} e^{-u^2/2} du ; \quad (2.5)$$

e.g., see Section VIII.4 of Feller (1971) and Chapter 7 of Chung (1974).

It is useful to review what the limit (2.4) means: The convergence in distribution means that the cdf's converge, i.e.,

$$P(n^{-1/2}(S_n - mn) \leq x) \rightarrow P(N(0, \sigma^2) \leq x) \quad \text{as } n \rightarrow \infty \quad (2.6)$$

for all x . More generally, given real-valued random variables Z_n , $n \geq 1$, and Z , there is convergence in distribution, by the standard definition, denoted by $Z_n \Rightarrow Z$, if the associated cdf's converge, i.e., if

$$F_n(x) \equiv P(Z_n \leq x) \rightarrow P(Z \leq x) \equiv F(x) \quad \text{as } n \rightarrow \infty \quad (2.7)$$

for all x that are continuity points of the limiting cdf F , i.e., for which $P(Z = x) = 0$.

Since the normal distribution has a continuous cdf, the restriction to continuity points of the limiting cdf in (2.7) does not arise in (2.6). We need to allow non-convergence at discontinuity points in (2.7), because we want to say that we have convergence $Z_n \Rightarrow Z$ in situations such as the special

case in which $P(Z = z) = 1$ and $P(Z_n = z_n) = 1$ for all n and $z_n \rightarrow z$ as $n \rightarrow \infty$. If $z_n \rightarrow z$ with $z_n > z$ for all n , then $F_n(z) \equiv P(Z_n \leq z) = 0$ for all n , while $F(z) \equiv P(Z \leq z) = 1$. Since $F_n(x) \rightarrow F(x)$ for all x except $x = z$, we obtain the desired convergence $Z_n \Rightarrow Z$ if we require pointwise convergence of the cdf's everywhere except at discontinuity points of the limiting cdf F .

There also are other convenient equivalent characterizations of convergence in distribution. In particular, (2.7) holds if and only if

$$E[h(Z_n)] \rightarrow E[h(Z)] \quad \text{as } n \rightarrow \infty \quad (2.8)$$

for every continuous bounded real-valued function h on $\mathbb{R}$, where E is the expectation operator. Moreover, (2.7) and (2.8) hold if and only if

$$g(Z_n) \Rightarrow g(Z) \quad \text{as } n \rightarrow \infty \quad (2.9)$$

for every continuous function g on $\mathbb{R}$. The alternative characterizations (2.8) and (2.9) are useful because they generalize to random elements of more general spaces.

The CLT in (2.4) explains the statistical regularity associated with the final positions of the centered random walks: In agreement with Figures 1.5 – 1.7, the CLT tells us that the centered partial sums $S_n - mn$ should be approximately normally distributed with mean 0 for all n sufficiently large.

We can also apply the CLT to obtain a corresponding limit for the scaled random walk $\mathbf{S}_n$ in (2.2) at an arbitrary time t in the interval $[0, 1]$. More generally, we can consider an arbitrary $t \geq 0$. To do so, we set $c_n = \sqrt{n}$ and $m = 1/2$. In particular, it is an easy consequence of (2.4) that we must have

$$n^{-1/2}(S_{[nt]} - m[nt]) \Rightarrow \sigma N(0, t) \quad \text{in } \mathbb{R} \quad \text{as } n \rightarrow \infty \quad (2.10)$$

for each $t \geq 0$, where $m = 1/2$ and $\sigma^2 = 1/12$.

From (2.10) we clearly see that the space-scaling constants c_n in (2.2) must be asymptotically equivalent to $c\sqrt{n}$ for some constant c as $n \rightarrow \infty$. Moreover, the space scaling by $\sqrt{n}$ is consistent with the units on the axes in Figures 1.2–1.4. Indeed, if we instead scale by $c_n = n^p$ for $p > 1/2$, then the values converge to 0 as $n \rightarrow \infty$. Similarly, if we scale by $c_n = n^p$ for $p < 1/2$, then the values diverge as $n \rightarrow \infty$. (The absolute values diverge to infinity.) This property can be confirmed by further analysis of simulations, but we do not pursue it.

We now want to convert (2.10) into a stochastic-process limit of the form (2.3). Note that the left side of (2.10) coincides with $\mathbf{S}_n(t)$, but the right side of (2.10) is not a stochastic process evaluated at time t . What we need to do is identify the appropriate limit process $\mathbf{S}$ in (2.3).

1.2.3. Identifying the Limit Process

We should recognize that we have arrived at another critical point. Another important intellectual step is needed here. *We not only must identify the limit process; we need to realize that there indeed should be a limit process.*

The appropriate limit process turns out to be a *Brownian motion* (BM). Brownian motion stochastic processes can be characterized as the real-valued stochastic processes with stationary and independent increments having continuous sample paths. Brownian motion evaluated at time t turns out to be normally distributed with mean mt and variance $\sigma^2 t$ for some constants m and σ^2 .

The special Brownian motion with parameters $m = 0$ and $\sigma^2 = 1$ is called *standard Brownian motion*; we shall refer to it by $\mathbf{B} \equiv \{\mathbf{B}(t) : t \geq 0\}$. It has marginal distributions

$$\mathbf{B}(t) \stackrel{d}{=} N(0, t), \quad t \geq 0, \quad (2.11)$$

where $\stackrel{d}{=}$ denotes equality in distribution.

An increment of Brownian motion is $\mathbf{B}(u) - \mathbf{B}(t)$ for $u > t$. By *stationary and independent increments*, we mean that the k -dimensional random vector

$$(\mathbf{B}(u_1 + h) - \mathbf{B}(t_1 + h), \dots, \mathbf{B}(u_k + h) - \mathbf{B}(t_k + h))$$

has a distribution independent of h for all k , and that the k component random variables are independent, providing that $0 \leq t_1 \leq u_1 \leq t_2 \leq \dots \leq u_k$.

Combining (2.10) and (2.11), we see that we can also express the limit (2.10) in terms of Brownian motion. In particular, after letting $c_n = \sqrt{n}$ in (2.2), we see that (2.10) is equivalent to

$$\mathbf{S}_n(t) \Rightarrow \sigma \mathbf{B}(t) \quad \text{in } \mathbb{R} \quad \text{as } n \rightarrow \infty \quad \text{for all } t \geq 0, \quad (2.12)$$

where $\mathbf{B}$ is a standard Brownian motion,

$$\mathbf{S}_n(t) \equiv n^{-1/2}(S_{\lfloor nt \rfloor} - m(\lfloor nt \rfloor)), \quad t \geq 0, \quad (2.13)$$

and $\sigma^2 = 1/12$ because the steps in the random walk are uniformly distributed over $[0, 1]$. In equations (2.11), (2.12) and (2.13) we have let t range over the semi-infinite interval $[0, \infty)$, but we could also have restricted t to the closed interval $[0, 1]$ to be consistent with the plots.

We can apply the limit in (2.12) to generate approximations for the terms of the original random walk. To generate approximations, we replace the convergence in distribution by approximate equality in distribution. From (2.12), we obtain the approximation

$$S_{[nt]} \approx m[nt] + n^{1/2}\sigma\mathbf{B}(t) \quad (2.14)$$

or

$$S_k \approx mk + n^{1/2}\sigma\mathbf{B}(k/n) , \quad (2.15)$$

where k is understood to be of order n and $\approx$ means approximately equal to in distribution. Note that the quality of the approximation for large n tends to depend more on the time scaling by n and the space scaling by $\sqrt{n}$ than the limit process $\sigma\mathbf{B}$.

The limit in (2.12) (with t ranging over the unit interval $[0, 1]$) can be regarded as the explanation for what we have seen in the random-walk plots. The limit in (2.12) is a *stochastic-process limit*, because it establishes convergence of the sequence of stochastic processes $\{\{\mathbf{S}_n(t) : 0 \leq t \leq 1\} : n \geq 1\}$ in (2.13) to the limiting stochastic process $\{\sigma\mathbf{B}(t) : 0 \leq t \leq 1\}$. However, we want to go beyond the limit as expressed via (2.12). We want to strengthen the form of convergence in order to be able to deduce convergence of related quantities of interest; in particular, we want to show that plots of the centered random walk converge to plots of standard Brownian motion as $n \rightarrow \infty$.

The probability law or distribution of a stochastic process is usually specified by the family of its finite-dimensional distributions (f.d.d.'s). Hence, a natural first step is to go beyond convergence of the one-dimensional marginal distributions, which is provided by (2.12), to convergence of the f.d.d.'s, i.e., the k -dimensional marginal distributions for all k . From the assumed independence among the random walk steps, it is not difficult to see that (2.12) can be extended to obtain

$$(\mathbf{S}_n(t_1), \dots, \mathbf{S}_n(t_k)) \Rightarrow (\sigma\mathbf{B}(t_1), \dots, \sigma\mathbf{B}(t_k)) \quad \text{in } \mathbb{R}^k \quad (2.16)$$

as $n \rightarrow \infty$ for all positive integers k and all k time points $t_1, \dots, t_k$ with $0 \leq t_1 < \dots < t_k \leq 1$, where convergence in distribution of random elements of $\mathbb{R}^k$ is defined by the natural generalization of (2.7), (2.8) or (2.9). Because of the independence among the random walk steps in this example, there is little difference between (2.12) and (2.16), but in general (2.16) is a much stronger conclusion.

However, we want to go even further. We want to go beyond convergence of the f.d.d.'s in (2.16) to convergence of the plots. We want to establish

limits for more general functions of the stochastic processes. To do so, *we regard $\mathbf{S}_n$ and $\mathbf{B}$ as random elements of a function space containing all possible sample paths.* (A function space is a space of functions.)

For $\mathbf{B}$, we could consider the space $C \equiv C([0, 1], \mathbb{R})$ of all continuous real-valued functions on the unit interval $[0, 1]$, but to include $\mathbf{S}_n$, we need discontinuous functions. (We could work with the space C if we used linearly interpolated random walks, as in (2.1), but we are considering the step functions.) We could consider a space containing all continuous functions and the special step functions that capture the structure of $\mathbf{S}_n$, but with other applications in mind, we consider a larger set of functions. We let the function space be the set $D \equiv D([0, 1], \mathbb{R})$ of all real-valued functions on $[0, 1]$ that are right-continuous at all t in $[0, 1)$ and have left limits everywhere in $(0, 1]$, endowed with an appropriate topology (notion of convergence, see Chapter 3).

The desired generalization of (2.12) and (2.16) follows from *Donsker's theorem*. Donsker's theorem is a *functional central limit theorem* (FCLT), which implies here that

$$\mathbf{S}_n \Rightarrow \sigma \mathbf{B} \quad \text{in } D, \quad (2.17)$$

where again $\mathbf{S}_n$ is the scaled random walk in (2.13), $\mathbf{B}$ is standard Brownian motion and the function space D is endowed with an appropriate topology. We discuss the topology on D and the precise meaning of (2.17) in Section 3.3.

Even though Brownian motion has a relatively simple characterization, it is a special stochastic process. For example, it has the self-similarity property observed in the plots (without limit). In particular, for all $c > 0$, the stochastic process $\{c^{-1/2}\mathbf{B}(ct) : 0 \leq t \leq 1\}$ has the same probability law on D ; equivalently, it has the same finite-dimensional distributions, i.e., the random vector $(c^{-1/2}\mathbf{B}(ct_1), \dots, c^{-1/2}\mathbf{B}(ct_k))$ has a distribution in $\mathbb{R}^k$ that is independent of c for any positive integer k and any k time points $t_i, 1 \leq i \leq k$, with $0 < t_1 < \dots < t_k \leq 1$.

Indeed, the self-similarity is a direct consequence of the stochastic-process limit in (2.17): First observe from (2.13) that, for any $c > 0$,

$$\mathbf{S}_{cn}(t) = c^{-1/2}\mathbf{S}_n(ct), \quad t \geq 0. \quad (2.18)$$

By taking limits on both sides of (2.18), we obtain

$$\{\mathbf{B}(t) : 0 \leq t \leq 1\} \stackrel{d}{=} \{c^{-1/2}\mathbf{B}(ct) : 0 \leq t \leq 1\}. \quad (2.19)$$

For further discussion, see Section 4.2.

Even though we are postponing a detailed discussion of the meaning of the convergence in (2.17), we can state a convenient characterization, which explains the applied value of (2.17) compared to (2.12) and (2.16). Just as in (2.8), the limit (2.17) means that

$$E[h(\mathbf{S}_n)] \rightarrow E[h(\sigma\mathbf{B})] \quad \text{as } n \rightarrow \infty \quad (2.20)$$

for every continuous bounded real-valued function h on D . The topology on D enters in by determining which functions h are continuous. Just as with (2.9), (2.20) holds if and only if

$$g(\mathbf{S}_n) \Rightarrow g(\sigma\mathbf{B}) \quad \text{in } \mathbb{R} \quad (2.21)$$

for every continuous real-valued function g on D . (It is easy to see that (2.20) implies (2.21) because the composition function $h \circ g$ is a bounded continuous real-valued function whenever g is continuous and h is a bounded continuous real-valued function.) Interestingly, (2.21) is the way that Donsker (1951) originally expressed his FCLT. The convergence of the functionals (real-valued functions) in (2.21) explains why the limit in (2.17) is called a FCLT.

It turns out that we also obtain (2.21) for every continuous function g , regardless of the range. For example, the function g could map D into D . Then we can obtain new stochastic-process limits from any given one. That is an example of the continuous-mapping approach for obtaining stochastic-process limits; see Section 3.4. The representation (2.21) is appealing because it exposes the applied value of (2.17) as an extension of (2.12) and (2.16). We obtain many associated limits from (2.21).

1.2.4. Limits for the Plots

We illustrate the continuous-mapping approach by establishing a limit for the plotted random walks, where as before we regard the plot as being in the unit square $[0, 1] \times [0, 1]$.

To establish limits for the plotted random walks, we use the functions $sup : D \rightarrow \mathbb{R}$, $inf : D \rightarrow \mathbb{R}$, $range : D \rightarrow \mathbb{R}$ and $plot : D \rightarrow D$, defined for any $x \in D$ by

$$sup(x) \equiv \sup_{0 \leq t \leq 1} x(t),$$

$$inf(x) \equiv \inf_{0 \leq t \leq 1} x(t),$$

$$range(x) \equiv sup(x) - inf(x)$$

and

$$plot(x) \equiv (x - inf(x))/range(x) .$$

Note that $plot(x)$ is an element of D for each $x \in D$ such that $range(x) \neq 0$. Moreover, the function $plot$ is scale invariant, i.e., for each positive scalar c and $x \in D$ with $range(x) \neq 0$,

$$plot(cx) = plot(x) .$$

Fortunately, these functions turn out to preserve convergence in the topologies we consider. (The first three functions are continuous, while the final $plot$ function is continuous at all x for which $range(x) \neq 0$, which turns out to be sufficient.) Hence we obtain the initial limits

$$n^{-1/2} \max_{1 \leq k \leq n} \{S_k - mk\} = sup(\mathbf{S}_n) \Rightarrow sup(\sigma \mathbf{B}) \equiv \sup_{0 \leq t \leq 1} \{\sigma \mathbf{B}(t)\} ,$$

$$n^{-1/2} \min_{1 \leq k \leq n} \{S_k - mk\} = inf(\mathbf{S}_n) \Rightarrow inf(\sigma \mathbf{B}) \equiv \inf_{0 \leq t \leq 1} \{\sigma \mathbf{B}(t)\} ,$$

$$n^{-1/2} range(\{S_k - mk : 0 \leq k \leq n\}) \equiv range(\mathbf{S}_n) \Rightarrow range(\sigma \mathbf{B})$$

in $\mathbb{R}$ and the final desired limit

$$plot(\mathbf{S}_n) \Rightarrow plot(\sigma \mathbf{B}) = plot(\mathbf{B}) \quad \text{in } D ,$$

where

$$plot(\{S_k - mk : 0 \leq k \leq n\}) = plot(\{c_n^{-1}(S_k - mk) : 0 \leq k \leq n\}) \equiv plot(\mathbf{S}_n) ,$$

from Donsker's theorem ((2.17) and (2.21)).

The limit $plot(\mathbf{S}_n) \Rightarrow plot(\mathbf{B})$ states that the plot of the scaled random walk converges to the plot of standard Brownian motion. Note that we use $plot$, not only as a function mapping D into D , but as a function mapping $\mathbb{R}^{n+1}$ into D taking the random walk segment into its plot.) Hence Donsker's theorem implies that the random walk plots can indeed be regarded as approximate plots of Brownian motion for all sufficiently large n . By using the FCLT refinement, we see that the stochastic-process limits do indeed explain the statistical regularity observed in the plots.

To highlight this important result, we state it formally as a theorem. Later chapters will provide a proof; specifically, we can apply Sections 3.4, 12.7 and 13.4.

Theorem 1.2.1. (convergence of plots to the plot of standard Brownian motion) *Consider an arbitrary stochastic sequence $\{S_k : k \geq 0\}$. Suppose that the limit in (2.3) holds in the space D with one of the Skorohod non-uniform topologies, where $c_n = \sqrt{n}$ and $\mathbf{S} = \sigma \mathbf{B}$ for some positive constant σ , with $\mathbf{B}$ being standard Brownian motion, as occurs in Donsker's theorem. Then*

$$\text{plot}(\{S_k - mk : 0 \leq k \leq n\}) \Rightarrow \text{plot}(\mathbf{B}) .$$

But an even more general result holds: *We have convergence of the plots for any space-scaling constants and almost any limit process.* We have the following more general theorem (proved in the same way as Theorem 1.2.1).

Theorem 1.2.2. (convergence of plots associated with any stochastic-process limit) *Consider an arbitrary stochastic sequence $\{S_k : k \geq 0\}$. Suppose that the limit in (2.3) holds in the space D with one of the Skorohod non-uniform topologies, where c_n and $\mathbf{S}$ are arbitrary. If*

$$P(\text{range}(\mathbf{S}) = 0) = 0 ,$$

then

$$\text{plot}(\{S_k - mk : 0 \leq k \leq n\}) \Rightarrow \text{plot}(\mathbf{S}) .$$

Note that the functions *sup*, *inf*, *range* and *plot* depend on more than one value $x(t)$ of the function x ; they depend on the function over an initial segment. Thus, we exploit the strength of the limit in D in (2.17) as opposed to the limit in $\mathbb{R}$ in (2.12) or even the limit in $\mathbb{R}^k$ in (2.16). For the random walk we have considered (with IID uniform random steps), the three forms of convergence in (2.12), (2.16) and (2.17) all hold, but in general (2.16) is strictly stronger than (2.12) and (2.17) is strictly stronger than (2.16). Formulating the stochastic-process limits in D means that we can obtain many more limits for related quantities of interest, because many more quantities of interest can be represented as images of continuous functions on the space of stochastic-process sample paths.

Remark 1.2.2. *Limits for the relative final position.* As noted in Remark 1.1.1, if we look at the final position of the centered random walk in the plots, ignoring the units on the axes, then we actually see the relative final position of the centered random walk, as defined in (1.1). Statistical regularity for the relative final position also follows directly from Theorems 1.2.1 and 1.2.2, because the relative final position is just the plot evaluated at time 1, i.e.,

$plot(x)(1)$. Provided that 1 is almost surely a continuity point of the limit process $\mathbf{S}$, under the conditions of Theorem 1.2.2 we have

$$R_n \Rightarrow plot(\mathbf{S})(1) \quad \text{in } \mathbb{R} \quad \text{as } n \rightarrow \infty,$$

as a consequence of the continuous-mapping approach, using the projection map that maps $x \in D$ into $x(1)$. ■

To summarize, the random-walk plots *reveal* remarkable statistical regularity associated with large n because the plotter automatically does the required scaling. In turn, the stochastic-process limits *explain* the statistical regularity observed in the plots. In particular, Donsker's FCLT implies that the random-walk plots converge in distribution to the plots of standard Brownian motion as $n \rightarrow \infty$.

1.3. Invariance Principles

The random walks we have considered so far are very special: the steps are IID with a uniform distribution in the interval $[0, 1]$. However, the great power of the SLLN, FSLLN, CLT and FCLT is that they hold much more generally. Essentially the same limits hold in many situations in which the step distribution is changed or the IID condition is relaxed, or both. Moreover, the limits each depend on only a single parameter of the random walk. The limits in the SLLN and the FSLLN only involve the single parameter m , which is the mean step size in the IID case. Similarly, after centering is done, the limits in the CLT and FCLT only involve the single parameter σ^2 , which is the variance of the step size in the IID case. Thus these limit theorems are *invariance principles*.

Moreover, the plots have an even stronger invariance property, because the limiting plots have no parameters at all! (We are thinking of the plot being in the unit square $[0, 1] \times [0, 1]$ in every case, ignoring the units on the axes.) Assuming only that the mean is positive, the plots of the uncentered random walk (with arbitrary step-size distribution) approach the identity function $e \equiv e(t) \equiv t$, $0 \leq t \leq 1$. If instead the mean is negative, then the limiting plot is $-e$ over the interval $[0, 1]$. Similarly, the plots of the centered random walks approach the plot of standard Brownian motion over $[0, 1]$; i.e., the limiting plot does not depend on the variance σ^2 . Thus, the random-walk plots reveal remarkable statistical regularity!

The power of the invariance principles is phenomenal. We will give some indication by giving a few examples and by indicating how they can be

applied. We recommend further experimentation to become a true believer. For example, the plots of the partial sums – centered and uncentered – should be contrasted with corresponding plots for the random-walk steps. Even for large n , plots of uniform random numbers and exponential (exponentially distributed) random numbers look very different, whereas the plots of the corresponding partial sums look the same (for all n sufficiently large).

1.3.1. The Range of Brownian Motion

We can apply the invariance property to help determine limiting probability distributions. For example, we can apply the invariance property to help determine the distribution of the limiting random variables $\sup(\mathbf{B})$ and $\text{range}(\mathbf{B})$.

We first consider the supremum $\sup(\mathbf{B})$. We can use combinatorial methods to calculate the distribution of $\max_{1 \leq k \leq n} \{S_k - km\}$ for any given n for the special case of the *simple random walk*, with $P(X_1 = +1) = P(X_1 = -1) = 1/2$, as shown in Chapter III of Feller (1968) or Section 11 of Billingsley (1968). In that way, we obtain

$$P(\sup(\mathbf{B}) > x) = 2P(N(0, 1) > x) \equiv 2\Phi^c(x) , \quad (3.1)$$

where $\Phi^c(t) \equiv 1 - \Phi(t)$ for Φ in (2.5). Since $\sup(\mathbf{B}) \stackrel{d}{=} |N(0, 1)|$,

$$E[\sup(\mathbf{B})] = \sqrt{2/\pi} \approx 0.8 \quad (3.2)$$

and

$$E[\sup(\mathbf{B})^2] = E[N(0, 1)^2] = 1 .$$

These calculations are not entirely elementary; for details see 26.2.3, 26.2.41 and 26.2.46 in Abramowitz and Stegun (1972).

The limit $\text{range}(\sigma\mathbf{B})$ is more complicated, but it too can be characterized; see Section 11 of Billingsley (1968) and Borodin and Salminen (1996). There the combinatorial methods for the simple random walk are used again to determine the joint distribution of $\inf(\mathbf{B})$ and $\sup(\mathbf{B})$, yielding

$$P(a < \inf(\mathbf{B}) < \sup(\mathbf{B}) < b) = \sum_{k=-\infty}^{k=+\infty} (-1)^k [\Phi(b+k(b-a)) - \Phi(a+k(b-a))] ,$$

where Φ is again the standard normal cdf. From (3.2), we see that the mean of the range is

$$E[\text{range}(\mathbf{B})] = E[\sup(\mathbf{B})] - E[\inf(\mathbf{B})] = 2E[\sup(\mathbf{B})] = 2\sqrt{2/\pi} \approx 1.6.$$

We can perform multiple replications of random-walk simulations to estimate the distribution of $\text{range}(\mathbf{B})$ and associated summary characteristics such as the variance. We show the estimate of the probability density function of $\text{range}(\mathbf{B})$ based on 10,000 samples of the random walk with 10,000 steps, each uniformly distributed on $[0, 1]$, in Figure 1.14 (again obtained using the nonparametric density estimator *density* from *S*). The range of the centered random walk should be approximately $\sigma\sqrt{n}$ times the range $\text{range}(\mathbf{B})$, so we divide the observed ranges in this experiment by $\sqrt{n/12} = 28.8675$. The estimated mean and standard deviation of $\text{range}(\mathbf{B})$ were 1.58 and 0.474, respectively. The estimated 0.1, 0.25, 0.5, 0.75 and 0.9 quantiles were 1.05, 1.24, 1.50, 1.85 and 2.23, respectively. This characterization of the distribution of $\text{range}(\mathbf{B})$ helps us interpret what we see in the random-walk plots.

From the analysis above, we know approximately what the mean and standard deviation of the range should be in the random-walk plots. Since $E[\text{range}(\mathbf{B})] \approx 1.6$, the mean of the random walk range should be about $1.6\sigma\sqrt{n} \approx 0.46\sqrt{n}$. Similarly, since the standard deviation of $\text{range}(\mathbf{B})$ is approximately 0.47, the standard deviation of the range in the random-walk plot should be approximately $0.47\sigma\sqrt{n} \approx 0.14\sqrt{n}$. Hence the (mean, standard deviation) pairs in Figures 1.3 and 1.4 with $n = 10^4$ and $n = 10^6$ are, respectively, (46, 14) and (460, 140). Note that the six observed values in each case are consistent with these pairs.

Historically, the development of the limiting behavior of $\sup(\mathbf{S}_n)$ played a key role in the development of the general theory; e.g. see the papers by Erdős and Kac (1946), Donsker (1951), Prohorov (1956) and Skorohod (1956). ■

Remark 1.3.1. Fixed space scaling. In our plots, we have let the plotter automatically determine the units on the vertical axis. Theorems 1.2.1 and 1.2.2 show that there is striking statistical regularity associated with automatic plotting. However, for comparison, it is often desirable to have common units. Interestingly, Donsker's FCLT and the analysis of the range above shows how to determine appropriate units for the vertical axis for the centered random walk, before the simulations are run.

First, the CLT and FCLT tell us the range of values for the centered random walk should be of order $\sqrt{n}$ as the sample size n grows. The invariance principle tells us that, for suitably large n the scaling should depend on the random-walk-step distribution only through its variance σ^2 .

The limit for the supremum $\sup(\mathbf{S}_n) \equiv n^{-1/2} \max_{1 \leq k \leq n} \{S_k - mk\}$ tells us more precisely what fixed space scaling should be appropriate for the

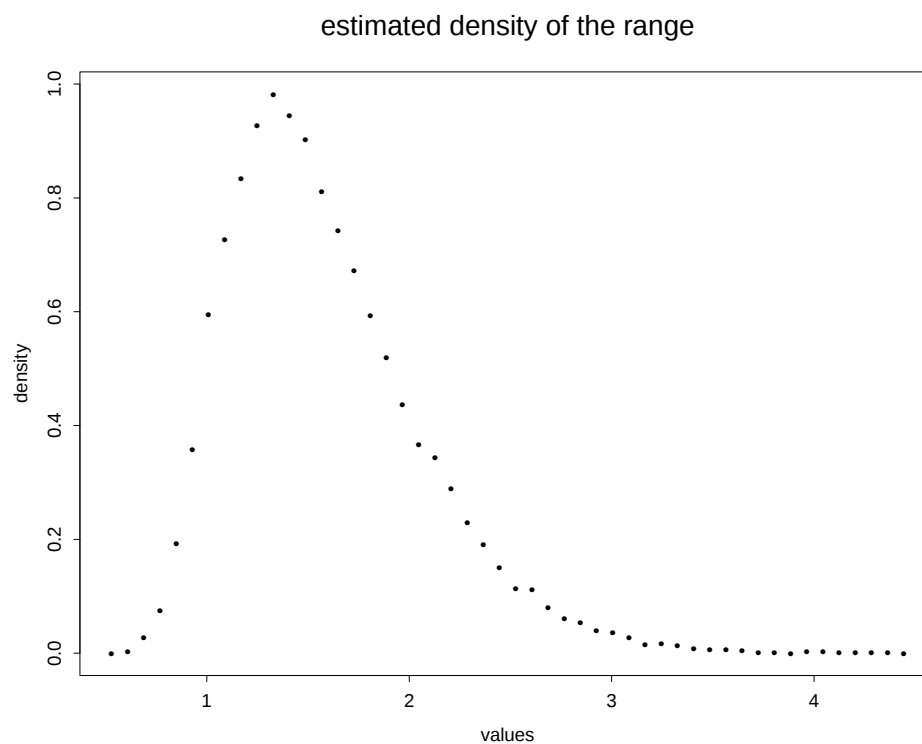


Figure 1.14: An estimate of the probability density of the range of Brownian motion over $[0, 1]$, obtained from 10,000 independent samples of random walks with 10,000 steps, each step being uniformly distributed in the interval $[0, 1]$.

plots. Since $2P(N(0,1) \geq 4)$ may be judged suitably small, from (3.1) we conclude that it should usually be appropriate to let the values on the vertical axis for a centered random walk fall in the interval $[-4\sigma\sqrt{n}, 4\sigma\sqrt{n}]$ as a function of n and σ^2 . For example, we could use this space scaling to replot the six random-walk plots in Figure 1.4. Since $n = 10^6$ and $\sigma^2 = 1/12$ there, we would let the values on the vertical axes in Figure 1.4 fall in the interval $[-1155, 1155]$. Notice that the values for the six plots all fall in the interval $[-700, 450]$, so that this fixed space scaling would work in Figure 1.4. ■

To gain a better appreciation of the invariance property, we perform some more simulations. First, we want to see that the IID conditions are *not* necessary.

1.3.2. Relaxing the IID Conditions

To illustrate how the IID conditions can be relaxed, we consider *exponential smoothing*.

Example 1.3.1. Exponential smoothing. We now consider a simple example of a random walk in which the steps are neither independent nor identically distributed. We let the steps be constructed by exponential smoothing. Equivalently, the steps are an autoregressive moving-average (ARMA) process of order (1,0); see Section 4.6.

In particular, suppose that we generate uniform random numbers U_k on the interval $[0, 1]$, $k \geq 1$, as before, but we now let the k^{th} step of the random walk be defined recursively by

$$X_k \equiv (1 - \gamma)X_{k-1} + \gamma U_k, \quad k \geq 1, \quad (3.3)$$

where $X_0 = U_0$, where U_0 is another uniform random number on $[0, 1]$ and $0 < \gamma < 1$. Clearly, the new random variables X_k are neither independent nor identically distributed. Moreover, the distribution of X_k is no longer uniform. It is not difficult to see, though, that as k increases the distribution of X_k approaches a nondegenerate limit. More generally, the sequence $\{X_{n+k} : k \geq 0\}$ is asymptotically stationary as $n \rightarrow \infty$, but successive random variables remain dependent.

We now regard the random variables X_k as steps of a random walk; i.e., we let the successive positions of the random walk be

$$S_k \equiv X_1 + \cdots + X_k, \quad k \geq 1, \quad (3.4)$$

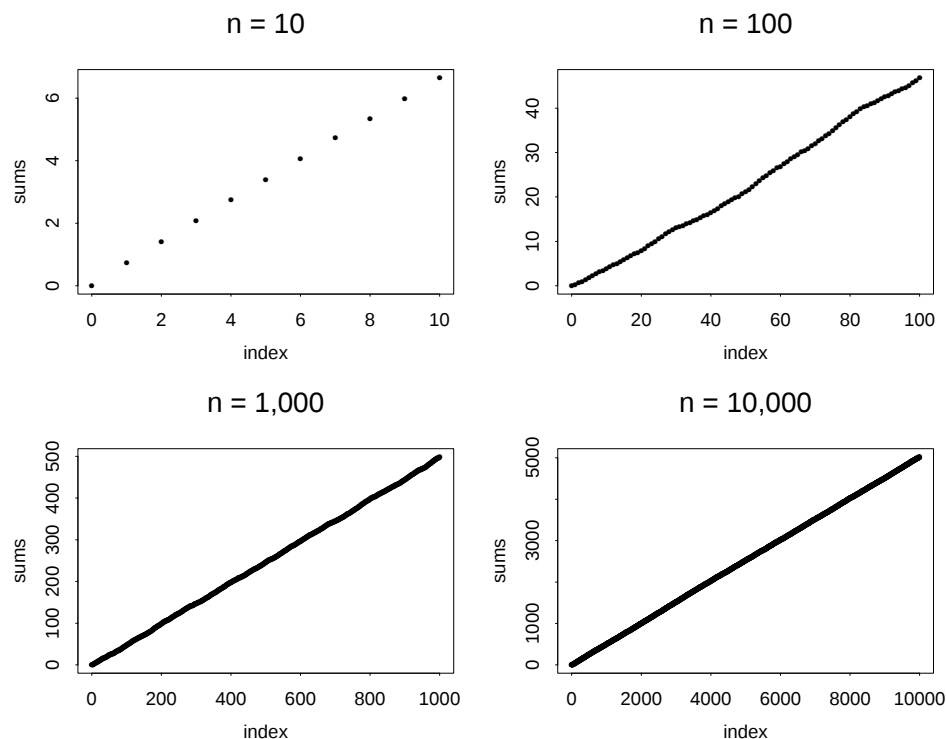


Figure 1.15: Possible realizations of the first 10^j steps of the uncentered random walk $\{S_k : k \geq 0\}$ with steps constructed by exponential smoothing, as in (3.3), for $j = 1, \dots, 4$.

where $S_0 \equiv 0$. Next we repeat the experiments done before. We display plots of the uncentered and centered random walks with $\gamma = 0.2$ for $n = 10^j$ with $j = 1, \dots, 4$ in Figures 1.15 and 1.16. To determine the appropriate centering constant (the steady-state mean of X_k), we solve the equation

$$E[X] = (1 - \gamma)E[X] + \gamma E[U]$$

to obtain $m \equiv E[X] = E[U] = 1/2$. Even though the distribution of X_k changes with k , the mean remains unchanged because of our choice of the initial condition.

Figures 1.15 and 1.16 look much like Figures 1.1 and 1.2 for the IID case. However, there is some significant difference for small n because the successive steps are positively correlated, causing the initial steps to be alike. However, the plots look like the previous plots for larger n . For the

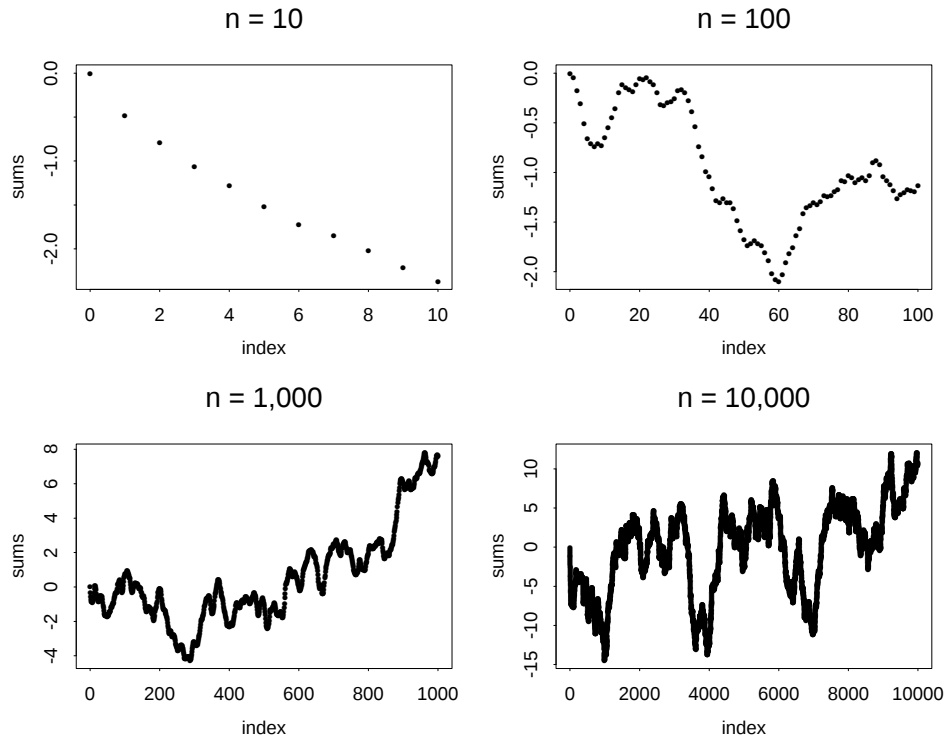


Figure 1.16: Possible realizations of the first 10^j steps of the centered random walk $\{S_k - k/2 : k \geq 0\}$ with steps constructed by exponential smoothing, as in (3.3), for $j = 1, \dots, 4$.

centered random walks in Figure 1.16 with $n = 10^4$, what we see is again approximately a plot of Brownian motion. ■

We can easily construct many other examples of random walks with dependent steps. For instance, we could consider a *random walk in a random environment*. A simple example has a two-state Markov-chain environment process with transition probabilities $P_{1,2} = 1 - P_{1,1} = p$ and $P_{2,1} = 1 - P_{2,2} = q$ for $0 < p < 1$ and $0 < q < 1$. We then let the k^{th} step X_k have one distribution if the Markov chain is in state 1 at the k^{th} step, and another distribution if the Markov chain is in state 2 then. We first run the Markov chain. Then, conditional on the realized states of the Markov chain, the random variables X_k are mutually independent with the appropriate distributions (depending upon the state of the Markov chain). If we consider a stationary version of the Markov chain, then the sequence $\{X_k : k \geq 1\}$ is stationary. Regardless of the initial conditions, we again see the same statistical regularity in the associated partial sums when n is sufficiently large. We invite the reader to consider such examples.

1.3.3. Different Step Distributions

Now let us return to random walks with IID steps and consider different possible step distributions. We now repeat the experiments above with various functions of the uniform random numbers, i.e., for $X_k \equiv f(U_k)$, $1 \leq k \leq n$, for different real-valued functions f . In particular, consider the following three cases:

$$\begin{aligned} \text{(i)} \quad X_k &\equiv -m \log(1 - U_k) \quad \text{for } m = 1, 10 \\ \text{(ii)} \quad X_k &\equiv U_k^p \quad \text{for } p = 1/2, 3/2 \\ \text{(iii)} \quad X_k &\equiv U_k^{-1/p} \quad \text{for } p = 1/2, 3/2 . \end{aligned} \tag{3.5}$$

As before, we form partial sums associated with the new summands X_k , just as in (3.4).

Before actually considering the plots, we observe that what we are doing covers the general IID case. Given the sequence of IID random variables $\{U_k : k \geq 1\}$, by the method above we can create an associated sequence of IID random variables $\{X_k : k \geq 1\}$ where X_k has an arbitrary cdf F . Letting the left-continuous inverse of F be

$$F^{\leftarrow}(t) \equiv \inf\{s : F(s) \geq t\}, \quad 0 < t < 1 ,$$

we can obtain the desired random variables X_k with cdf F by letting

$$X_k \equiv F^{\leftarrow}(U_k), \quad k \geq 1. \quad (3.6)$$

Since

$$F^{\leftarrow}(s) \leq t \quad \text{if and only if} \quad F(t) \geq s, \quad (3.7)$$

we obtain

$$P(F^{\leftarrow}(U) \leq t) = P(U \leq F(t)) = F(t),$$

where U is a random variable uniformly distributed on $[0, 1]$, which implies that $F^{\leftarrow}(U)$ has cdf F for any cdf F when U is uniformly distributed on $[0, 1]$. For example, we see that X_k has an exponential distribution with mean m in case (i) of (3.5): If $F(t) = e^{-t/m}$, then $F^{\leftarrow}(t) = -m \log(1 - t)$ and

$$P(X_k > t) = P(-m \log(1 - U_k) > t) = P(1 - U_k < e^{-t/m}) = e^{-t/m}.$$

Incidentally, we could also work with the right-continuous inverse of F , defined by

$$F^{-1}(t) \equiv \inf\{s : F(s) > t\} = F^{\leftarrow}(t+), \quad 0 < t < 1,$$

where $F^{\leftarrow}(t+)$ is the right limit at t , because

$$P(F^{-1}(U) = F^{\leftarrow}(U)) = 1,$$

since $F^{\leftarrow}$ and F^{-1} differ at, at most, countably many points.

Moreover, $F^{\leftarrow}(U_k)$, $k \geq 1$, are IID when U_k , $k \geq 1$, are IID. Of course, there also are other ways to generate IID random variables with specified distributions, but what we are doing is often a natural way.

So let us plot the uncentered and centered random walks with the step sizes in (3.5). When we do so for cases (i) and (ii), we see essentially the same pictures as before. For example, plots of the first 10^4 steps of the centered random walks in the four cases in (i) and (ii) of (3.5) are shown in Figure 1.17.

Again the plots look like plots of Brownian motion, indistinguishable from the plots for the uniform steps in Figure 1.3. Note that the units on the y axis change from plot to plot, but the plots themselves tend to have a common distribution.

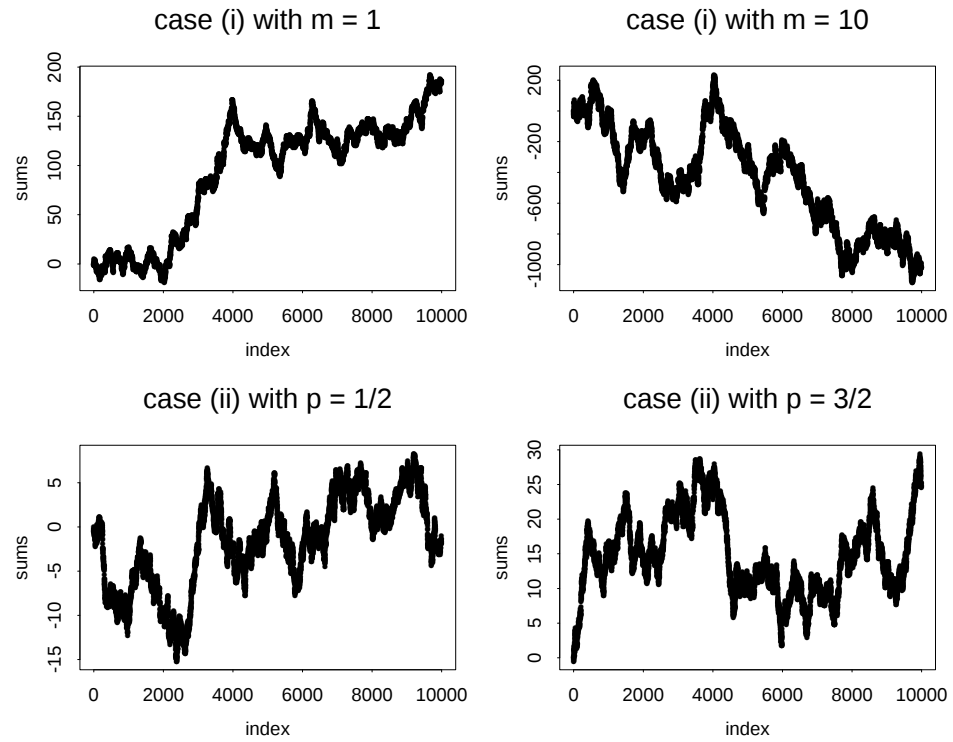


Figure 1.17: Possible realizations of the first 10^4 steps of the random walk $\{S_k - mk : k \geq 0\}$ with steps distributed as X_k in cases (i) and (ii) of (3.5).

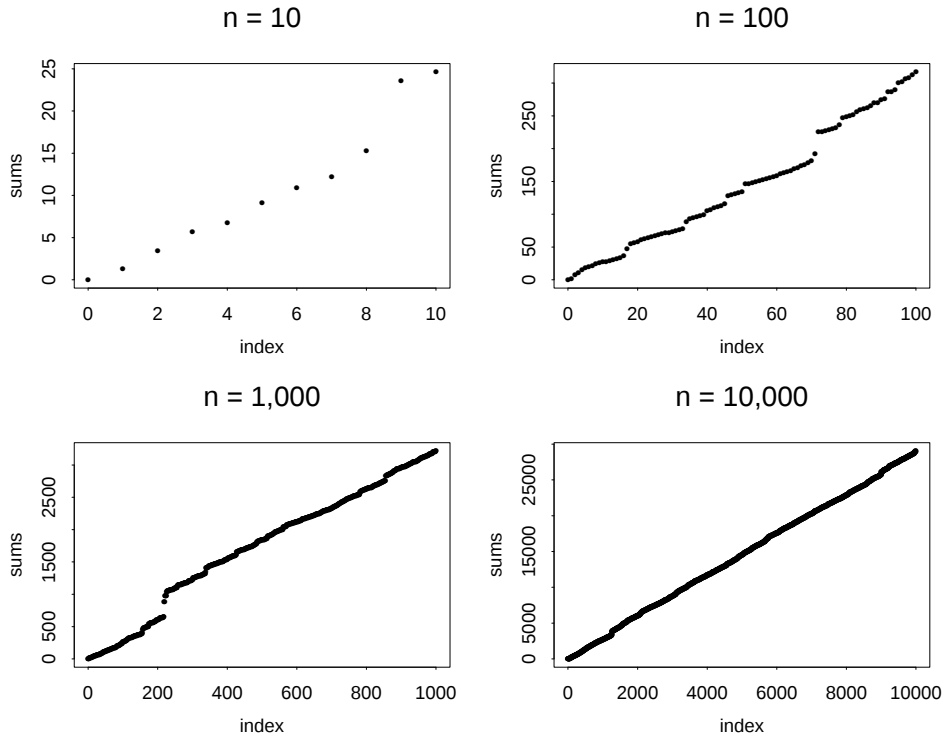


Figure 1.18: Possible realizations of the first 10^j steps of the uncentered random walk $\{S_k : k \geq 0\}$ with steps distributed as $U_k^{-1/p}$ in case (iii) of (3.5) for $p = 3/2$ and $j = 1, \dots, 4$.

1.4. The Exception Makes the Rule

Just when boredom has begun to set in, after seeing the same thing in cases (i) and (ii) in (3.5), we should be ready to appreciate the startlingly different large- n pictures in case (iii). Plots of the uncentered random walks are plotted in Figures 1.18 and 1.19.

In the case $p = 3/2$ in Figure 1.18, the plot of the uncentered random walk is again approaching a line as $n \rightarrow \infty$, but not as rapidly as before. (Again we ignore the units on the axes when we look at the plots.) However, in the case $p = 1/2$ in Figure 1.19 we see something radically different: For large n , the plots have *jumps*!

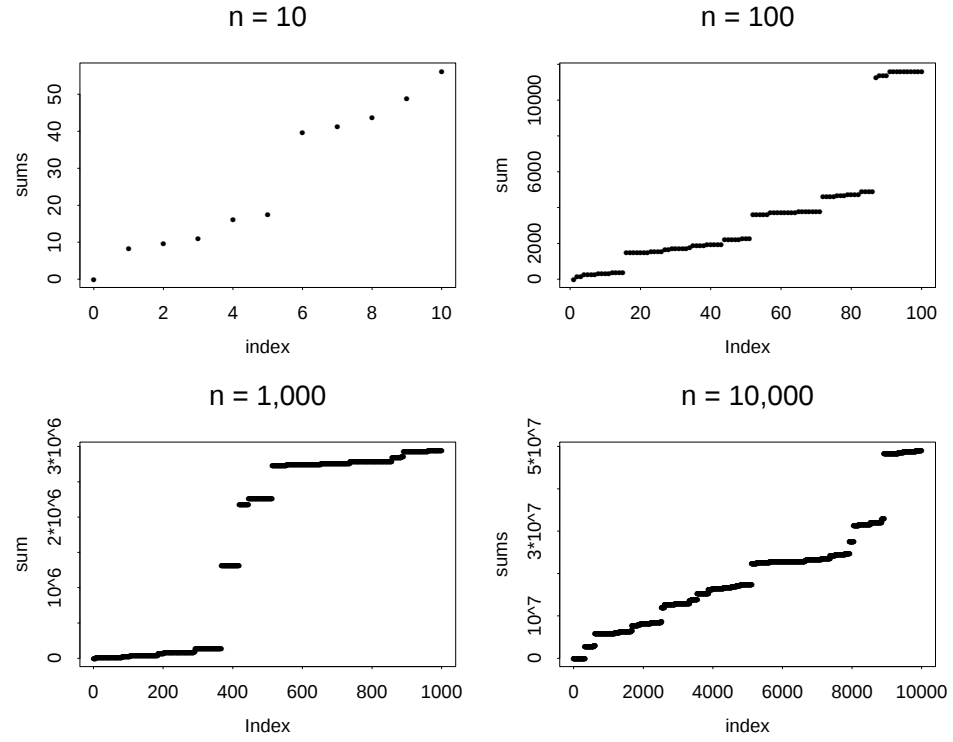


Figure 1.19: Possible realizations of the first 10^j steps of the uncentered random walk $\{S_k : k \geq 0\}$ with steps distributed as $U_k^{-1/p}$ in case (iii) of (3.5) for $p = 1/2$ and $j = 1, \dots, 4$.

1.4.1. Explaining the Irregularity

Fortunately, probability theory again provides an explanation for the *irregularity* that we now see: The SLLN states, under the prevailing IID assumptions, that scaled partial sums $n^{-1}S_n$ will approach the mean EX_1 w.p.1 as $n \rightarrow \infty$, regardless of other properties of the probability distribution of X_1 , *provided that a finite mean exists*. Knowing the SLLN, we should expect to see lines when $n = 10^4$ in all experiments except possibly in case (iii).

We might initially be fooled in case (iii), but we should anticipate occasional large steps because $U^{-1/p}$ involves *dividing* by very small values when U is small. Upon more careful examination, we see that $U^{-1/p}$ has a *Pareto distribution* with parameter p , which we refer to as $\text{Pareto}(p)$, when U is uniformly distributed on $[0, 1]$, i.e.,

$$P(U^{-1/p} > t) = P(U < t^{-p}) = t^{-p}, \quad t \geq 1, \quad (4.1)$$

with mean

$$E(U^{-1/p}) = \int_0^\infty P(U^{-1/p} > t) dt = 1 + \int_1^\infty t^{-p} dt, \quad (4.2)$$

which is finite, and equal to $1 + (p - 1)^{-1}$, if and only if $p > 1$; see Chapter 19 of Johnson and Kotz (1970) for background on the Pareto distribution and Lemma 1 on p. 150 of Feller (1971) for the integral representation of the mean.

Thus the SLLN tells us not to expect the same behavior observed in the previous experiments in case (iii) when $p \leq 1$. Thus, unlike all previous random walks considered, the conditions of the SLLN are *not satisfied* in case (iii) with $p = 1/2$.

Now let us consider the random walk with $\text{Pareto}(p)$ steps for $p = 3/2$ in (3.5) (iii). Consistent with the SLLN, Figure 1.18 shows that the plots are approaching a straight line as $n \rightarrow \infty$ in this case. But what happens when we center?

1.4.2. The Centered Random Walk with $p = 3/2$

So now let us consider the centered random walk in case (iii) with $p = 3/2$. (Since the mean is infinite when $p = 1/2$, we cannot center when $p = 1/2$. We will return to the case $p = 1/2$ later.) We center by subtracting the mean, which in the case $p = 3/2$ is $1 + (p - 1)^{-1} = 3$. Plots of the centered random walk with $p = 3/2$ for $n = 10^j$ with $j = 1, 2, 3, 4$ are shown in Figure

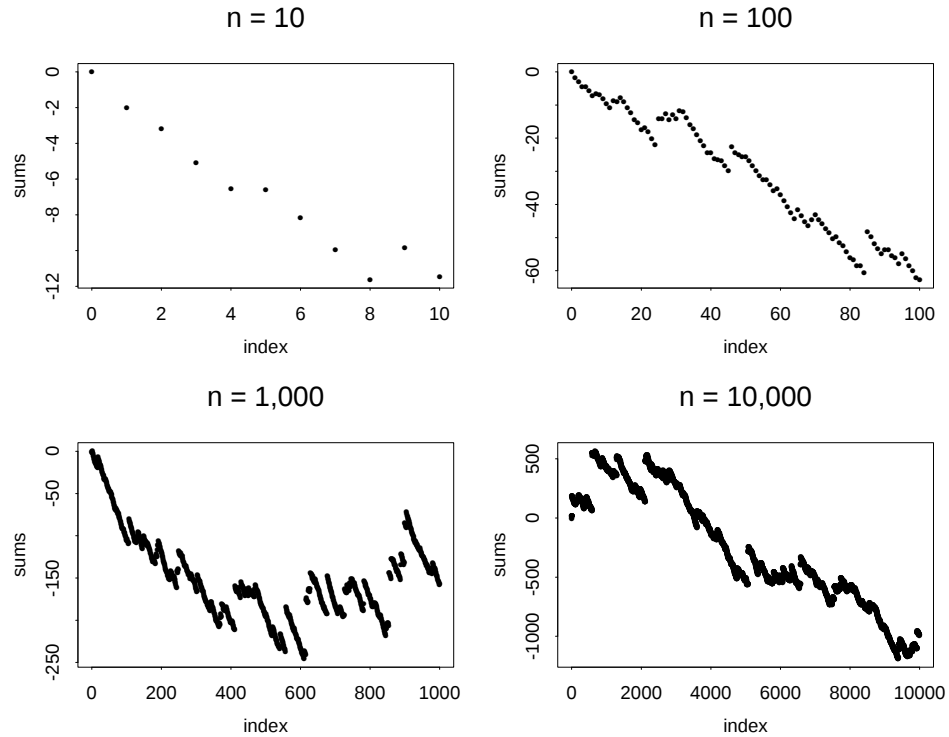


Figure 1.20: Possible realizations of the first 10^j steps of the centered random walk $\{S_k - 3k : k \geq 0\}$ associated with the Pareto steps $U_k^{-1/p}$ for $p = 3/2$, having mean 3 and infinite variance, for the cases $j = 1, \dots, 4$.

1.20. As before, the centering causes the plotter to automatically blow up the picture. However, now the slight departures from linearity for large n in Figure 1.18 are magnified. Now, just as in Figure 1.19, we see jumps in the plot!

Once again, probability theory offers an explanation. Just as the SLLN ceases to apply when the IID summands have infinite mean, so does the (classical) CLT cease to apply when the IID summands have finite mean but infinite variance. Such a case occurs with the Pareto(p) summands in case (iii) in (3.5) when $1 < p \leq 2$. Thus, consistent with what we see in Figure 1.18, the SLLN holds, but the CLT does not, for the Pareto(p) random variable $U^{-1/p}$ in case (iii) when $p = 3/2$.

We have arrived at another critical point, where an important intellectual step is needed. We need to recognize that, *even though the sample paths are*

very different from the previous random-walk plots, which are approaching plots of Brownian motion, there may still be important statistical regularity in the new plots with jumps.

To see the statistical regularity, we need to repeat the experiment and consider larger values of n . Even though the plots look quite different from the previous random-walk plots, we can see statistical regularity in the plots (again ignoring the units on the axes). To confirm that observation, six possible realizations for $p = 3/2$ in the cases $n = 10^4$ and $n = 10^6$ are shown in Figures 1.21 and 1.22. Figures 1.21 and 1.22 show more irregular paths, but with their own distinct character, much like handwriting. (We might contemplate the probability of the path writing a word. With a suitable font for the script, we might see “Null” but not “Set”.) Again, Figures 1.21 and 1.22 show that there is statistical regularity associated with the irregularity we see. The plots are independent of n for all n sufficiently large. Again we see self-similarity in the plots.

Even though the irregular paths in Figures 1.19 – 1.22 have jumps, as before we can look for statistical regularity through the distribution of these random paths. Again, to be able to see something, we can focus on the final positions. Focusing first on the case with $p = 3/2$, we plot the estimated density of the centered sums $S_n - 3n$ for $n = 1,000$. Once again, we obtain the density estimate by performing independent replications of the experiment. To have more data this time, we use 10,000 independent replications. We display the resulting density estimate in Figure 1.23.

When we look at the estimated density of the final position, we see that it is radically different from the previous density plots in Figures 1.5 and 1.7. Clearly, *the final position is no longer normally distributed!*

Nevertheless, there is statistical regularity. As before, when we repeat the experiment with different random number seeds, we obtain essentially the same result for all sufficiently large n . Examination shows that there is statistical regularity, just as before, but the approximating distribution of the final position is now different. In Figure 1.23, the peak of the density looks like a spike because the range of values is now much greater. In turn, the range of values is greater because the distribution of $S_n - 3n$ has a heavy tail.

The heavier tails are more clearly revealed when we plot the tail of the empirical cdf of the observed values. (By the tail of a cdf F , we mean the *complementary cdf* or *ccdf*, defined by $F^c(t) \equiv 1 - F(t)$.)

To focus on the tail of the cdf F , we plot the tail of the empirical cdf in $\log - \log$ scale in Figure 1.18; i.e., we plot $\log F^c(t)$ versus $\log t$. To use $\log - \log$ scale, we consider only those values greater than 1, of which there

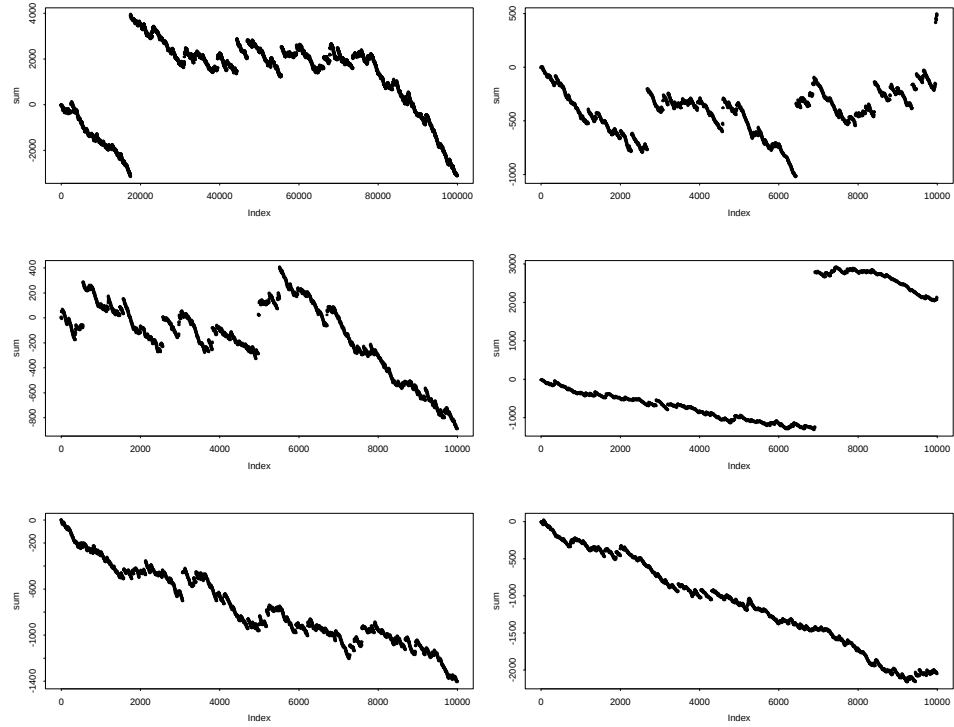


Figure 1.21: Six independent realizations of the first 10^4 steps of the centered random walk $\{S_k - 3k : k \geq 0\}$ associated with the Pareto steps $U_k^{-1/p}$ for $p = 3/2$, having mean 3 and infinite variance.

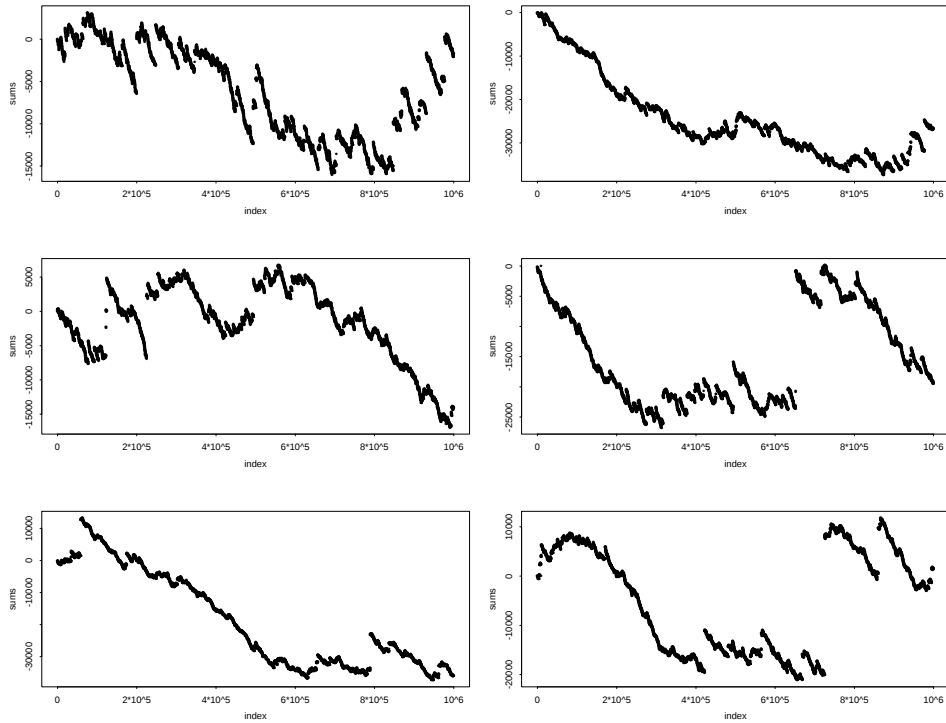


Figure 1.22: Six independent realizations of the first 10^6 steps of the centered random walk $\{S_k - 3k : k \geq 0\}$ associated with the Pareto steps $U_k^{-1/p}$ for $p = 3/2$, having mean 3 and infinite variance.

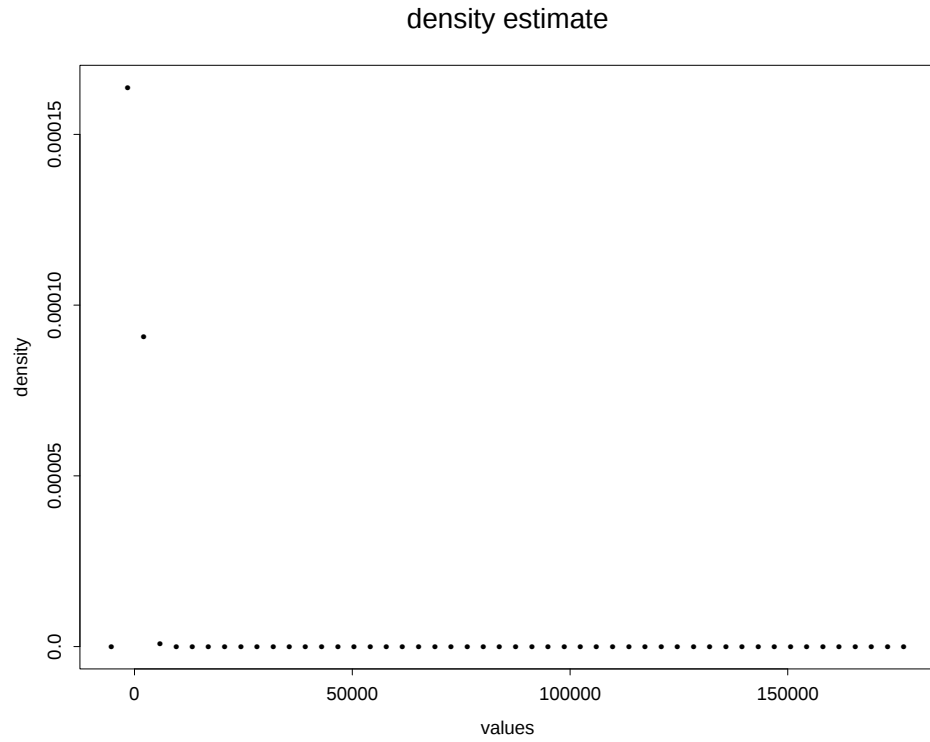


Figure 1.23: The density estimate obtained from 10,000 independent samples of the final position of the centered random walk (i.e., the centered partial sum $S_{1000} - 3000$) associated with the Pareto steps $U_k^{-1/p}$ for $p = 3/2$.

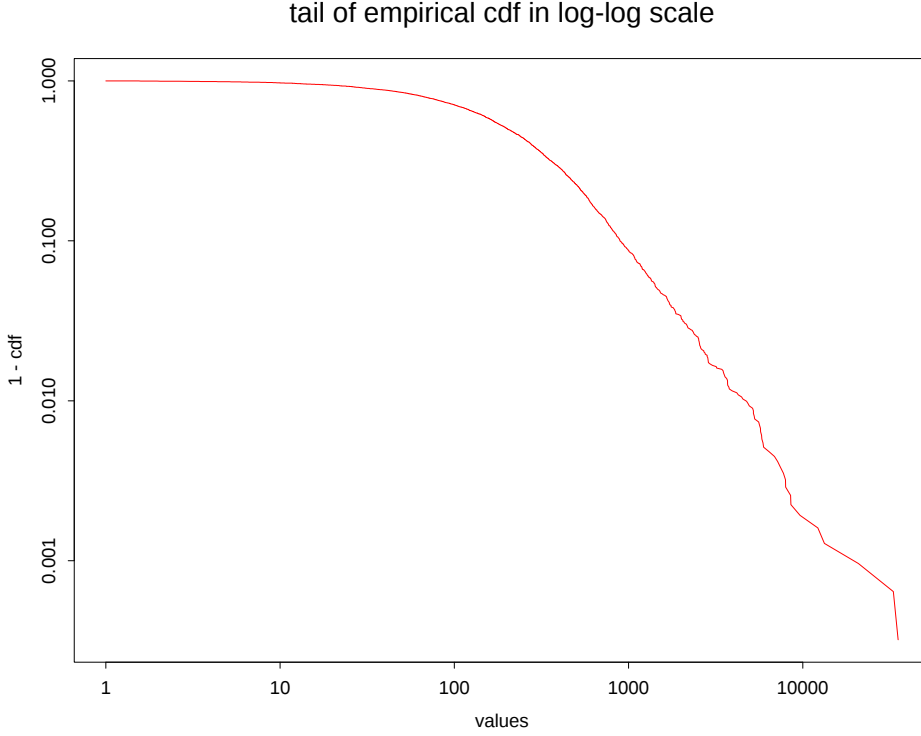


Figure 1.24: The tail of the empirical distribution function in $\log - \log$ scale obtained from 10,000 independent samples of the final position of the centered random walk (i.e., the partial sum $S_{1000} - 3000$) associated with the Pareto steps $U_k^{-1/p}$ for $p = 3/2$ corresponding to the density in Figure 1.23. The results are based on the 3,121 values greater than 1.

were 3,121 when $n = 10^4$.

From Figure 1.24, we see that for larger values of the argument t , the empirical cdf has a linear slope in $\log - \log$ scale. That indicates a *power tail*. Indeed, if the cdf is of the form

$$F^c(t) = \alpha t^{-\beta} \quad \text{for } t \geq t_0 > 1, \quad (4.3)$$

then

$$\log F^c(t) = -\beta \log t + \log \alpha \quad (4.4)$$

for $t > t_0$. Then the parameters α and β in (4.3) can be seen as the intercept and slope in the $\log - \log$ plot.

Again there is supporting theory: A generalization of the CLT implies, under the IID assumptions and other regularity conditions (satisfied here), that properly scaled versions of the centered partial sums of Pareto(p) random steps converge in distribution, as in (2.7). In particular, when $1 < p < 2$,

$$n^{-1/p}(S_n - mn) \Rightarrow L \quad \text{in } \mathbb{R}, \quad (4.5)$$

where $m = 1 + (p - 1)^{-1}$ is the mean and the limiting random variable L has a *non-Gaussian stable law* (depending upon p); e.g., see Chapter XVII of Feller (1971). In our specific case of $p = 3/2$, we have space scaling by $n^{2/3}$.

Unlike the Pareto distribution, the limiting stable law is not a pure power, but it has a power tail; i.e., it is asymptotically equivalent to a power: for $1 < p < 2$,

$$P(L > t) \sim ct^{-p} \quad \text{as } t \rightarrow \infty \quad (4.6)$$

for some positive constant c , where $f(t) \sim g(t)$ as $t \rightarrow \infty$ means that f is *asymptotically equivalent* to g , i.e., $f(t)/g(t) \rightarrow 1$ as $t \rightarrow \infty$. Thus the tail of the limiting stable law has the same asymptotic decay rate as the Pareto distribution of a single step.

Unlike the standard CLT in (2.4), the space scaling in (4.5) involves $c_n = n^{1/p}$ for $1 < p < 2$ instead of $c_n = n^{1/2}$. Nevertheless, the generalized CLT shows that there is again remarkable statistical regularity in the centered partial sums when the mean is finite and the variance is infinite. We again obtain essentially the same probability distribution for all n . We also obtain essentially the same probability distribution for other nonnegative step distributions, provided that they are centered by subtracting the finite mean, and that the step-size cdf $F^c(t)$ has the same asymptotic tail; i.e., we require that

$$F^c(t) \sim ct^{-p} \quad \text{as } t \rightarrow \infty \quad (4.7)$$

for some positive constant c .

As before, there is also an associated stochastic-process limit. A generalization of Donsker's theorem (the FCLT) implies that the sequence of scaled random walks with Pareto(p) steps having $1 < p < 2$ converges in distribution to a *stable Lévy motion* as $n \rightarrow \infty$ in D . Now

$$\mathbf{S}_n \Rightarrow \mathbf{S} \quad \text{in } D, \quad (4.8)$$

where

$$\mathbf{S}_n(t) \equiv n^{-1/p}(S_{[nt]} - m[nt]), \quad 0 \leq t \leq 1, \quad (4.9)$$

for $n \geq 1$, m is the mean and $\mathbf{S}$ is a stable Lévy motion. That is, the stochastic-process limit (2.3) holds for $\mathbf{S}_n$ in (2.2), but now with $c_n = n^{1/p}$ and the limit process $\mathbf{S}$ being stable Lévy motion instead of Brownian motion. Moreover, a variant of the previous invariance property holds here as well. For nonnegative random variables (the step sizes) satisfying (4.7), the limit process depends on its distribution only through the decay rate p and the single parameter c appearing in (4.7). We discuss this FCLT further in Chapter 4.

Since the random walk steps are IID, it is evident that the limiting stable Lévy motion must have stationary and independent increments, just like Brownian motion. However, the marginal distributions in $\mathbb{R}$ or $\mathbb{R}^k$ are non-normal stable laws instead of the normal laws. Moreover, the stable Lévy motion has the self-similarity property, just like Brownian motion, but now with a different scaling. Now, for any $c > 0$, the stochastic process $\{c^{-1/p}\mathbf{S}(ct) : 0 \leq t \leq 1\}$ has a probability law on D , and thus finite-dimensional distributions, that are independent of c . Indeed, the proof is just like the proof for Brownian motion in (2.18).

It is significant that the space scaling to achieve statistical regularity is different now. In (4.9) above, we divide by $n^{1/p}$ for $1 < p < 2$ instead of by $n^{1/2}$. Similarly, in the self-similarity of the stable Lévy motion, we multiply by $c^{-1/p}$ instead of $c^{-1/2}$. The new scaling can be confirmed by looking at the values on the y-axis in the plots of Figures 1.20–1.22.

Figures 1.20–1.22 show that, unlike Brownian motion, stable Lévy motion must have *discontinuous sample paths*. Hence, *we have a stochastic-process limit in which the limit process has jumps*. The desire to consider such stochastic-process limits is a primary reason for this book.

1.4.3. Back to the Uncentered Random Walk with $p = 1/2$

Now let us return to the first Pareto(p) example with $p = 1/2$. The plots in Figure 1.19 are so irregular that we might not suspect that there is any statistical regularity there. However, after seeing the statistical regularity in the case $p = 3/2$, we might well think about reconsidering the case $p = 1/2$.

As before, we investigate by making some more plots. We have noted that we cannot center because the mean is infinite. So let us make more plots of the uncentered random walk with $p = 1/2$. Thus, in Figure 1.25 we plot six independent realizations of the uncentered random walk with 10^4 Pareto(0.5) steps. Now, even though these plots are highly irregular, with a single jump sometimes dominating the entire plot, we see remarkable statistical regularity.

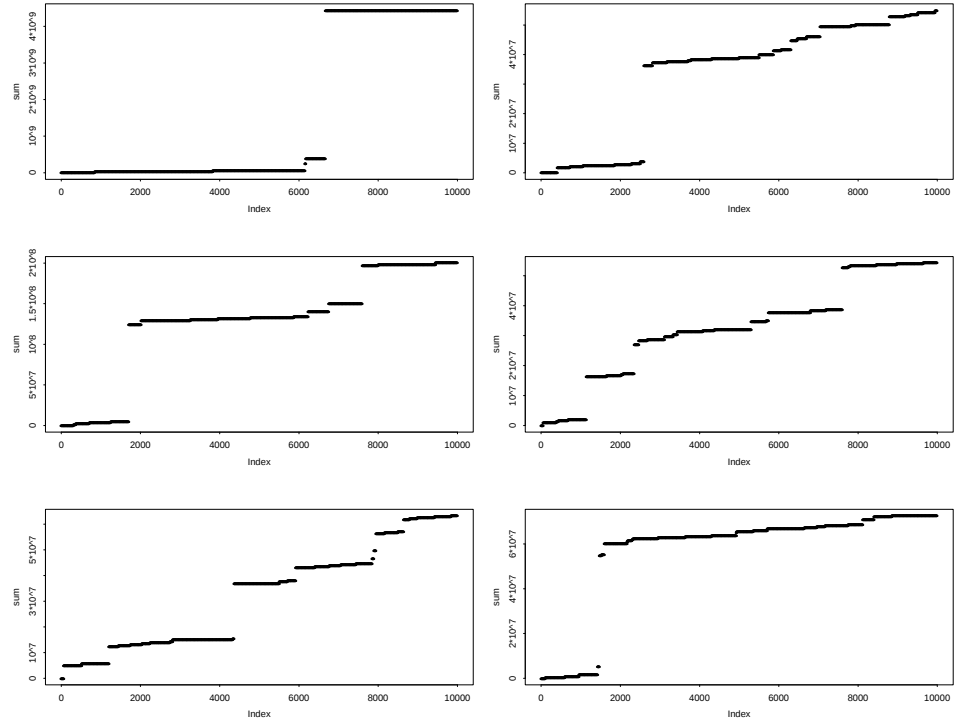


Figure 1.25: Six independent possible realizations of the first 10^4 steps of the uncentered random walk $\{S_k : k \geq 0\}$ with steps distributed as $U_k^{-1/p}$ in case (iii) of (3.5) for $p = 1/2$.

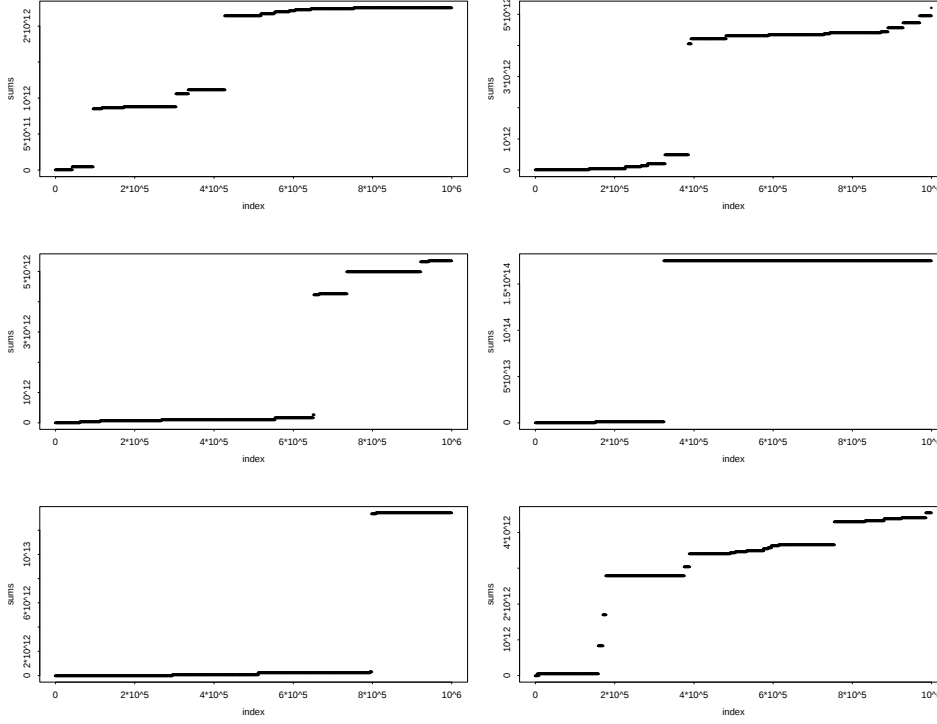


Figure 1.26: Six independent possible realizations of the first 10^6 steps of the uncentered random walk $\{S_k : k \geq 0\}$ with steps distributed as $U_k^{-1/p}$ in case (iii) of (3.5) for $p = 1/2$.

Paralleling Figures 1.4 and 1.22, we confirm what we see in Figure 1.25 by plotting six independent samples of the uncentered random walk in case (iii) with $p = 1/2$ for $n = 10^6$ in Figure 1.26. Even though the plots of the uncentered random walks with Pareto(0.5) steps in Figures 1.19 – 1.26 are radically different from the previous plots of centered and uncentered random walks, we see remarkable statistical regularity in the new plots. As before, the plots tend to be independent of n for all n sufficiently large, provided we ignore the units on the axes. Thus we see self-similarity, just as in the plots of the centered random walks before. *From the random-walk plots, we see that statistical regularity can occur in many different forms.*

Given what we have just done, it is natural to again look for statistical regularity in the final positions. Thus we consider the final positions S_n (without centering) for $n = 1000$ and perform 10,000 independent replica-

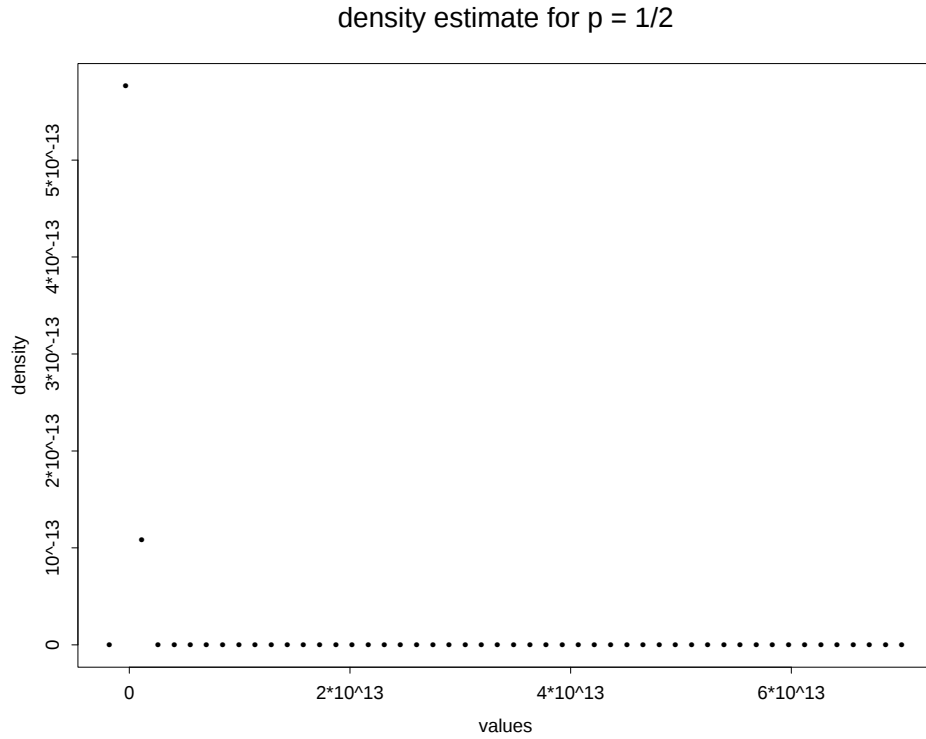


Figure 1.27: A density estimate obtained from 10,000 independent samples of the final position of the uncentered random walk (i.e., the partial sum S_{1000}) associated with the Pareto steps $U_k^{-1/p}$ in the case $p = 1/2$.

tions. Paralleling Figures 1.23 and 1.24 above, an estimate of the probability density and the tail of the empirical cdf are plotted in Figures 1.27 and 1.28 below.

Figures 1.27 and 1.28 are quite similar to Figures 1.23 and 1.24, but now the distribution has an even heavier tail. Again there is supporting theory: A generalization of the CLT states, under the IID assumptions and other regularity conditions (satisfied here), that for $0 < p < 1$ there is convergence in distribution of the *uncentered partial sums* to a non-Gaussian stable law if the partial sums are scaled appropriately, which requires that $c_n = n^{1/p}$. In particular, now with $p = 1/2$,

$$n^{-1/p} S_n \Rightarrow L \quad \text{in } \mathbb{R}, \quad (4.10)$$

where the limiting random variable again has a non-Gaussian stable law,

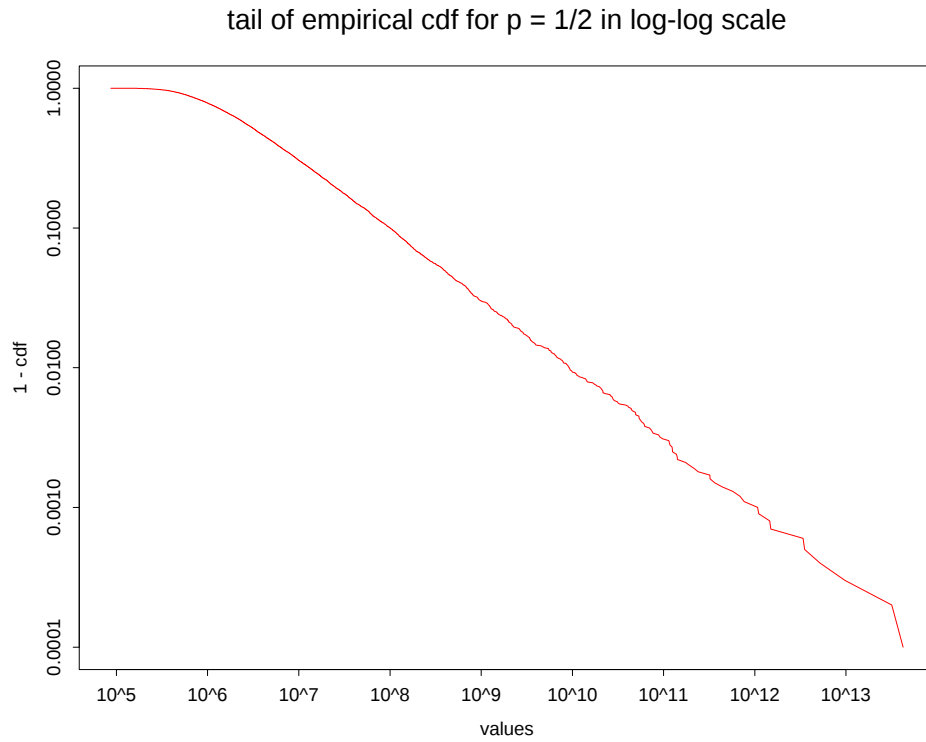


Figure 1.28: The tail of the empirical cumulative distribution function in *log-log* scale obtained from 10,000 independent samples of the final position of the uncentered random walk (i.e., the partial sum S_{1000}) associated with the Pareto steps $U_k^{-1/p}$ for $p = 1/2$ corresponding to the density in Figure 1.27.

which has an asymptotic power tail, i.e.,

$$P(L > t) \sim ct^{-p} \quad \text{as } t \rightarrow \infty \quad (4.11)$$

for $p = 1/2$ and some positive constant c ; again see Chapter XVII of Feller (1971). As before, the tail of the stable law has the same asymptotic decay rate as a single step of the random walk.

Moreover, there again is an associated stochastic-process limit. Another generalization of Donsker's FCLT implies that there is the stochastic-process limit (4.8), where

$$\mathbf{S}_n(t) \equiv n^{-1/p} S_{[nt]}, \quad 0 \leq t \leq 1, \quad (4.12)$$

for $n \geq 1$, with the limit process $\mathbf{S}$ being another stable Lévy motion depending upon p .

Again there is an invariance property: Paralleling (4.7), we require that the random-walk step ccdf F^c satisfy

$$F^c(t) \sim ct^{-p} \quad \text{as } t \rightarrow \infty, \quad (4.13)$$

where $p = 1/2$ and c is some positive constant. Any random walk with nonnegative (IID) steps having a ccdf satisfying (4.13) will satisfy the same FCLT, with the limit process depending on the step-size distribution only through the decay rate $p = 1/2$ and the constant c in (4.13).

As before, the plotter automatically does the proper scaling. However, the space scaling is different from both the previous two cases, now requiring division by $n^{1/p}$ for $p = 1/2$. Again, we can verify that the space scaling by $n^{1/p}$ is appropriate by looking at the values in the plots in Figures 1.19–1.26. Just as before, the stochastic-process limit in D implies that the limit process must be self-similar. Now, for any $c > 0$, the stochastic processes $\{c^{-1/p} \mathbf{S}(ct) : 0 \leq t \leq 1\}$ have probability laws in D that are independent of c .

Figures 1.19 and 1.25 show that the limiting stable Lévy motion for the case $p = 1/2$ must also have discontinuous sample paths. So we have yet another stochastic-process limit in which the limit process has jumps.

1.5. Summary

To summarize, in this chapter we have seen that there is remarkable statistical regularity associated with random walks as the number n of steps increases. That statistical regularity is directly revealed when we plot the

random walks. In great generality, as a consequence of Donsker's theorem, properly scaled versions of the centered random walks converge in distribution to Brownian motion as n increases. As a consequence, the random-walk plots converge to plots of standard Brownian motion.

The great generality of that result may make us forget that there are conditions for convergence to Brownian motion to hold. Through the exponential-smoothing example, we have seen that the conclusions of the classical limit theorems often still hold when the IID conditions are relaxed, but again there are limitations on the amount of dependence that can be allowed. That is easy to see by considering the extreme case in which *all the steps are identical!* Clearly, then the SLLN and the CLT break down. The classical limit theorems tend to remain valid when independence is replaced by *weak dependence*, but it is difficult to characterize the boundary exactly. We discuss FCLTs for weakly dependent sequences further in Chapter 4.

We also have seen for the case of IID steps that there are important situations in which the conditions of the FLLN and Donsker's FCLT do not hold. We have seen that these fundamental theorems are not valid in the IID case when the step-size distribution has infinite mean (the FLLN) or variance (the FCLT). Nevertheless, there often is remarkable statistical regularity associated with these heavy-tailed cases, but the limit process in the stochastic-process limit becomes a stable Lévy motion, which has jumps, i.e., it has discontinuous sample paths. We have thus seen examples of stochastic-process limits in which the limit process has jumps. We discuss such FCLTs further in Chapter 4.

If we allow greater dependence, which may well be appropriate in applications, then many more limit processes are possible, some of which will again have discontinuous sample paths. Again, see Chapter 4.

Chapter 2

Random Walks in Applications

The random walks we have considered in Chapter 1 are easy to think about, because they have a relatively simple structure. However, the random walks are abstract, so that they may seem disconnected from reality. But that is not so!

Even though the random walks are abstract, they play a fundamental role in many applications. Many stochastic processes in applied probability models are very closely related to random walks. Indeed, we are able to obtain many stochastic-process limits for stochastic processes of interest in applied probability models directly from established probability limits for random walks, using the continuous-mapping approach.

To elaborate on this important point, we now give three examples of stochastic processes closely related to random walks. The examples involve stock prices, the Kolmogorov-Smirnov test statistic and a queueing model for a buffer in a switch. In the final section we discuss the engineering significance of the queueing model and the (heavy-traffic) stochastic-process limits.

2.1. Stock Prices

In some applications, random walks apply very directly. A good example is finance, which often can be regarded as yet another game of chance; see *A Random Walk Down Wall Street* by Malkiel (1996).

Indeed, we might model the price of a stock over time as a random walk; i.e., the position S_n can be the price in time period n . However, it is common

to consider a refinement of the direct random-walk model, because the magnitude of any change is usually considered to be approximately proportional to the price.

A popular alternative model that captures that property is obtained by letting the price in period n be the *product* of the price in period $n - 1$ and a *random multiplier* Y_n ; i.e., if Z_n is the price in period n , then we have

$$Z_n = Z_{n-1}Y_n, \quad n \geq 1. \quad (1.1)$$

That in turn implies that

$$Z_n = Z_0(Y_1 \times \cdots \times Y_n), \quad n \geq 1. \quad (1.2)$$

Just as for random walks, for tractability we often assume that the successive random multipliers $Y_n : n \geq 1$, are IID. Hence, if we take logarithms, then we obtain

$$\log(Z_n) = \log(Z_0) + S_n, \quad n \geq 0,$$

where $\{S_n : n \geq 0\}$ is a random walk, defined as in (3.4), with steps $X_n \equiv \log(Y_n)$, $n \geq 1$ that are IID. With this multiplicative framework, the *logarithms of successive prices constitute an initial position plus a random walk*. Approximations for random walks thus produce direct approximations for the logarithms of the prices.

It is natural to consider limits for the stock prices, in which the duration of the discrete time periods decreases in the limit, so that we can obtain convergence of the sequence of discrete-time price processes to a continuous-time limit, representing the evolution of the stock price in continuous time. To do so, we need to change the random multipliers as we change n . We thus define a sequence of price models indexed by n . We let Z_k^n and Y_k^n denote the price and multiplier, respectively, in period k in model n . For each n , we assume that the sequence of multipliers $\{Y_k^n : k \geq 1\}$ is IID. Since the periods are shrinking as $n \rightarrow \infty$, we want $Y_k^n \rightarrow 1$ as $n \rightarrow \infty$. The general idea is to have

$$E[\log(Y_k^n)] \approx m/n \quad \text{and} \quad \text{Var}[\log(Y_k^n)] \approx \sigma^2/n.$$

We let the initial price be independent of n ; i.e., we let $Z_0^n \equiv Z_0$ for all n .

Thus, we incorporate the scaling within the partial sums for each n . We make further assumptions so that

$$\mathbf{S}_n(t) \equiv S_{[nt]}^n \Rightarrow \sigma \mathbf{B}(t) + mt \quad \text{as} \quad n \rightarrow \infty \quad (1.3)$$

for each $t > 0$, where $\mathbf{B}$ is standard Brownian motion. Given (1.3), we obtain

$$\log(\mathbf{Z}_n(t)) \equiv \log(Z_{[nt]}^n) = \log(Z_0) + S_{[nt]}^n \Rightarrow \log(Z_0) + \sigma \mathbf{B}(t) + mt ,$$

so that

$$\mathbf{Z}_n(t) \equiv Z_{[nt]}^n \Rightarrow \mathbf{Z}(t) \equiv Z_0 \exp(\sigma \mathbf{B}(t) + mt) ; \quad (1.4)$$

i.e., the price process converges in distribution as $n \rightarrow \infty$ to the stochastic process $\{\mathbf{Z}(t) : t \geq 0\}$, which is called *geometric Brownian motion*.

Geometric Brownian motion tends to inherit the tractability of Brownian motion. Since the moment generating function of a standard normal random variable is

$$\psi(\theta) \equiv E[\exp(\theta N(0, 1))] = \exp(\theta^2/2) ,$$

the k^{th} moment of geometric Brownian motion for any k can be expressed explicitly as

$$E[\mathbf{Z}(t)^k] = E[(Z_0)^k] \exp(kmt + k^2 t^2 \sigma^2 / 2) . \quad (1.5)$$

See Section 10.4 of Ross (1993) for an introduction to the application of geometric Brownian motion to finance, including a derivation of the Black-Scholes option pricing formula.

The analysis so far is based on the assumption that the random-walk steps $X_k^n \equiv \log(Y_k^n)$ are IID with finite mean and variance. However, even though the steps must be finite, the volatility of the stock market has led people to consider alternative models. If we drop the finite-mean or finite-variance assumption, then we can still obtain a suitable continuous-time approximation, but it is likely to be a geometric stable Lévy motion (obtained by replacing the Brownian motion by a stable Lévy motion in the exponential representation in (1.4)). Even other limits are possible when the steps come from a double sequence $\{\{X_k^n : k \geq 1\} : n \geq 1\}$. When we consider models for volatile prices, we should be ready to see stochastic-process limits with jumps. For further discussion, see Embrechts, Klüppelberg and Mikosch (1997), especially Section 7.6.

In addition to illustrating how random walks can be applied, this example illustrates that we sometimes need to consider double sequences of random variables, such as $\{\{X_k^n : k \geq 1\} : n \geq 1\}$, in order to obtain the stochastic-process limit we want.

2.2. The Kolmogorov-Smirnov Statistic

For our second random-walk application, let us return to the empirical cdf's considered in Example 1.1.1 in Section 1.1.3. What we want to see now is a stochastic-process limit for the difference between the empirical cdf and the underlying cdf, explaining the statistical regularity we saw in Figure 1.8. The appropriate limit process is the *Brownian bridge* $\mathbf{B}_0$, which is just Brownian motion $\mathbf{B}$ over the interval $[0, 1]$ conditioned to be 0 at the right endpoint $t = 1$.

Recall that the applied goal is to develop a statistical test to determine whether or not data from an unknown source can be regarded as an independent sample from a candidate cdf F . The idea is to base the test on the “difference” between the candidate cdf and the empirical cdf. We determine whether or not the observed difference is significantly greater than the difference for an independent sample from the candidate cdf F is likely to be. The problem, then, is to characterize the probability distribution of the difference between a cdf and the associated empirical cdf obtained from an independent sample. Interestingly, even here, random walks can play an important role.

Hence, let F be an arbitrary continuous candidate cdf and let F_n be the associated empirical cdf based on an independent sample of size n from F . A convenient test statistic, called the *Kolmogorov-Smirnov statistic*, can be based on the limit

$$D_n \equiv \sqrt{n} \sup_{t \in \mathbb{R}} \{ |F_n(t) - F(t)| \} \Rightarrow \sup(|\mathbf{B}_0|) \quad \text{as } n \rightarrow \infty, \quad (2.1)$$

where $\mathbf{B}_0$ is the Brownian bridge, which can be represented as

$$\mathbf{B}_0(t) = \mathbf{B}(t) - t\mathbf{B}(1), \quad 0 \leq t \leq 1, \quad (2.2)$$

$$\sup(|\mathbf{B}_0|) \equiv \sup_{0 \leq t \leq 1} \{ |\mathbf{B}_0(t)| \}$$

and

$$P(\sup(|\mathbf{B}_0|) > x) = 2 \sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2 x^2}, \quad x > 0. \quad (2.3)$$

Notice that the limit in (2.1) is independent of the cdf F (assuming only that the cdf F is continuous). The candidate cdf F could be the uniform cdf in Example 1.1.1, a normal cdf, a Pareto cdf or a stable cdf. In particular, the limit process here is unaffected by the cdf F having a heavy tail.

In practice, we would compute the Kolmogorov-Smirnov statistic D_n in (2.1) for the empirical cdf associated with the data from the unknown source and the candidate cdf F . We then compute, using (2.3), the approximate probability of observing a value as large or larger than the observed value of the Kolmogorov-Smirnov statistic, under the assumption that the empirical cdf does in fact come from an independent sample from F . If that probability is very small, then we would reject the hypothesis that the data come from an independent sample from F .

As usual, good judgement is needed in the interpretation of the statistical analysis. When the sample size n is not large, we might be unable to reject the hypothesis that the data is an independent sample from a cdf F for more than one candidate cdf F . On the other hand, with genuine data (not a simulation directly from the cdf F), for any candidate cdf F , we are likely to be able to reject the hypothesis that the data is an independent sample from F for all n sufficiently large. Our concern here, though, is to justify the limit (2.1).

So, how do random walks enter in? Random walks appear in two ways. First, the empirical cdf $F_n(t)$ as a function of n itself is a minor modification of a random walk. In particular,

$$nF_n(t) = \sum_{k=1}^n I_{(-\infty, t]}(X_k) ,$$

where $I_A(x)$ is the *indicator function* of the set A , with $I_A(x) = 1$ if $x \in A$ and $I_A(x) = 0$ otherwise. Thus, for each t , $nF_n(t)$ is the sum of the n IID Bernoulli random variables $I_{(-\infty, t]}(X_k)$, $1 \leq k \leq n$, and is thus a random walk.

Note that the Bernoulli random variable $I_{(-\infty, t]}(X_k)$ has mean $F(t)$ and variance $F(t)F^c(t)$. Hence we can apply the SLLN and the CLT to deduce that

$$F_n(t) \rightarrow F(t) \quad w.p.1 \quad \text{as } n \rightarrow \infty$$

and

$$\sqrt{n}(F_n(t) - F(t)) \Rightarrow N(0, F(t)F^c(t)) \quad \text{in } \mathbb{R} \quad \text{as } n \rightarrow \infty \quad (2.4)$$

for each $t \in \mathbb{R}$. Note that we have to multiply the difference by $\sqrt{n}$ in (2.4) in order to get a nondegenerate limit. That explains the multiplicative factor $\sqrt{n}$ in (2.1).

Paralleling the way we obtained stochastic-process limits for random walks in Section 1.2, we can go from the limit in (2.4) to the limit in (2.1)

by extending the limit in (2.4) to a stochastic-process limit in the function space D . We can establish the desired stochastic-process limit in D in two steps: first, by reducing the case of a general continuous cdf F to the case of the uniform cdf (i.e., the cdf of the uniform distribution on $[0, 1]$) and, second, by treating the case of the uniform cdf. Random walks can play a key role in the second step.

To carry out the first step, we show that the distribution of D_n in (2.1) is independent of the continuous cdf F . For that purpose, let $U_k, 1 \leq k \leq n$, be uniform random variables (on $[0, 1]$) and let G_n be the associated empirical cdf. Recall from equation (3.7) in Section 1.3.3 that

$$F^{\leftarrow}(U_k) \leq t \quad \text{if and only if} \quad U_k \leq F(t) ,$$

so that $F^{\leftarrow}(U_k) \stackrel{d}{=} X_k, 1 \leq k \leq n$, and

$$\{G_n(F(t)) : t \in \mathbb{R}\} \stackrel{d}{=} \{F_n(t) : t \in \mathbb{R}\} .$$

Hence,

$$D_n \equiv \sqrt{n} \sup_{t \in \mathbb{R}} \{|F_n(t) - F(t)|\} \stackrel{d}{=} \sqrt{n} \sup_{t \in \mathbb{R}} \{|G_n(F(t)) - F(t)|\} .$$

Moreover, since F is a continuous cdf, F maps $\mathbb{R}$ into the interval $(0, 1)$ plus possibly $\{0\}$ and $\{1\}$. Since $P(U = 0) = P(U = 1) = 0$ for a uniform random variable U , we have

$$D_n \stackrel{d}{=} \sqrt{n} \sup_{0 \leq t \leq 1} \{|G_n(t) - t|\} , \tag{2.5}$$

which of course is the special case for a uniform cdf.

Now we turn to the second step, carrying out the analysis for the special case of a uniform cdf, i.e., starting from (2.5). To make a connection to random walks, we exploit a well known property of Poisson processes. We start by focusing on the uniform order statistics: Let $U_k^{(n)}$ be the k^{th} order statistic associated with n IID uniform random variables; i.e., $U_k^{(n)}$ is the k^{th} smallest of the uniform random numbers. It is not difficult to see that the supremum in the expression for D_n in (2.5) must occur at one of the jumps in G_n (either the left or right limit) and these jumps occur at the random times $U_k^{(n)}$. Since each jump of D_n in (2.5) has magnitude $1/\sqrt{n}$,

$$|D_n - \sqrt{n}(\max_{1 \leq k \leq n} \{|U_k^{(n)} - k/n|\})| \leq 1/\sqrt{n} . \tag{2.6}$$

Now we can make the desired connection to random walks: It turns out that

$$(U_1^{(n)}, \dots, U_n^{(n)}) \stackrel{d}{=} (S_1/S_{n+1}, \dots, S_n/S_{n+1}) , \quad (2.7)$$

where

$$S_k \equiv X_1 + \dots + X_k, \quad 1 \leq k \leq n+1 ,$$

with $X_k, 1 \leq k \leq n+1$, being IID exponential random variables with mean 1. To justify relation (2.7), consider a Poisson process and let the k^{th} point be located at S_k (Which makes the intervals between points IID exponential random variables). It is well known, and easy to verify, that the first n points of the Poisson process are distributed in the interval $(0, S_{n+1})$ as the n uniform order statistics over the interval $(0, S_{n+1})$; e.g., see p. 223 of Ross (1993). When we divide by S_{n+1} we obtain the uniform order statistics over the interval $(0, 1)$, just as in the left side of (2.7).

With the connection to random walks established, we can apply Donsker's FCLT for the random walk $\{S_k : k \geq 0\}$ to establish the limit (2.1). In rough outline, here is the argument:

$$\begin{aligned} D_n &\approx \sqrt{n} \max_{1 \leq k \leq n} \{ | (S_k/S_{n+1}) - (k/n) | \} \\ &\approx (n/S_{n+1}) \max_{1 \leq k \leq n} \{ | (S_k - k)/\sqrt{n} - (k/n)(S_{n+1} - n)/\sqrt{n} | \} . \end{aligned} \quad (2.8)$$

Since $n/S_{n+1} \rightarrow 1$ as $n \rightarrow \infty$ and $(S_{n+1} - S_n)/\sqrt{n} \rightarrow 0$ as $n \rightarrow \infty$, we have

$$D_n \approx \sup_{0 \leq t \leq 1} \{ | (S_{[nt]} - [nt])/ \sqrt{n} - ([nt]/n)(S_n - n)/\sqrt{n} | \} . \quad (2.9)$$

To make the rough argument rigorous, and obtain (2.9), we repeatedly apply an important tool – the convergence-together theorem – which states that $X_n \Rightarrow X$ whenever $Y_n \Rightarrow X$ and $d(X_n, Y_n) \Rightarrow 0$, where d is an appropriate distance on the function space D ; see Theorem 11.4.7.

Since the functions $\psi_1 : D \rightarrow D$ and $\psi_2 : D \rightarrow \mathbb{R}$, defined by

$$\psi_1(x)(t) \equiv x(t) - tx(1), \quad 0 \leq t \leq 1 , \quad (2.10)$$

and

$$\psi_2(x) \equiv \sup_{0 \leq t \leq 1} \{ |x(t)| \} \quad (2.11)$$

are continuous, from (2.9) we obtain the desired limit

$$D_n \Rightarrow \sup_{0 \leq t \leq 1} \{ |B(t) - tB(1)| \} . \quad (2.12)$$

Finally, it is possible to show that relations (2.2) and (2.3) hold.

The argument here follows Breiman (1968, pp. 283–290). Details can be found there, in Karlin and Taylor (1980, p. 343) or in Billingsley (1968, pp. 64, 83, 103, 141). See Pollard (1984) and Shorack and Wellner (1986) for further development. See Borodin and Salminen (1996) for more properties of Brownian motion.

Historically, the derivation of the limit in (2.1) is important because it provided a major impetus for the development of the general theory of stochastic-process limits; see the papers by Doob (1949) and Donsker (1951, 1952), and subsequent books such as Billingsley (1968).

2.3. A Queueing Model for a Buffer in a Switch

Another important application of random walks is to queueing models. We will be exploiting the connection between random walks and queueing models throughout the queueing chapters. We only try to convey the main idea now.

To illustrate the connection between random walks and queues, we consider a discrete-time queueing model of data in a buffer of a switch or router in a packet communication network.

Let W_k represent the workload (or buffer content, which may be measured in bits) at the end of period k . During period k there is a random input V_k and a deterministic constant output μ (corresponding to the available bandwidth) provided that there is content to process or transmit. We assume that the successive inputs V_k are IID, although that is not strictly necessary to obtain the stochastic-process limits.

More formally, we assume that the successive workloads can be defined recursively by

$$W_k \equiv \min\{K, \max\{0, W_{k-1} + V_k - \mu\}\}, \quad k \geq 1, \quad (3.1)$$

where the initial workload is W_0 and the buffer capacity is K . The *maximum* appears in (3.1) because the workload is never allowed to become negative; the output (up to μ) occurs only when there is content to emit. The *minimum* appears in (3.1) because the workload is not allowed to exceed the capacity K at the end of any period; we assume that input that would make the workload exceed K at the end of the period is lost.

The workload process $\{W_k : k \geq 1\}$ specified by the recursion (3.1) is quite elementary. Since the inputs V_k are assumed to be IID, the stochastic process $\{W_k\}$ is a discrete-time Markov process. If, in addition, we assume

that the inputs V_k take values in a discrete set $\{ck : k \geq 0\}$ for some constant c (which is not a practical restriction), we can regard the stochastic process $\{W_k\}$ as a discrete-time Markov chain (DTMC). Since the state space of the DTMC $\{W_k\}$ is one-dimensional, the finite state space will usually not be prohibitively large. Thus, it is straightforward to exploit numerical methods for DTMC's, as in Kemeny and Snell (1960) and Stewart (1994), to describe the behavior of the workload process.

Nevertheless, we are interested in establishing stochastic-process limits for the workload process. In the present context, we are interested in seeing how the distribution of the inputs V_k affects the workload process. We can use heavy-traffic stochastic-process limit to produce simple formulas describing the performance. (We start giving the details in Chapter 5.) Those simple formulas provide insight that can be gained only with difficulty from a numerical algorithm for Markov chains.

We also are interested in the heavy-traffic stochastic-process limits to illustrate what can be done more generally. The heavy-traffic stochastic-process limits can be established for more complicated models, for which exact performance analysis is difficult, if not impossible. Since the heavy-traffic stochastic-process limits strip away unessential details, they reveal the key features determining the performance of the queueing system.

Now we want to see the statistical regularity associated with the workload process for large n . We could just plot the workload process for various candidate input processes $\{V_k : k \geq 1\}$ and parameters K and μ . However, the situation here is more complicated than for the random walks we considered previously. We can simply plot the workload process and let the plotter automatically do the scaling for us, but it is not possible to automatically see the desired statistical regularity. For the queueing model, we need to do some analysis to determine how to do the proper scaling in order to achieve the desired statistical regularity. (That is worth verifying.)

2.3.1. Deriving the Proper Scaling

It turns out that stochastic-process limits for the workload process are intimately related to stochastic-process limits for the random walk $\{S_k : k \geq 0\}$ with steps

$$X_k \equiv V_k - \mu ,$$

but notice that in general this random walk is not centered. The random walk is only centered in the special case in which the input rate $E[V_k]$ exactly matches the potential output rate μ . However, to have a well-behaved

system, we want the long-run potential output rate to exceed the long-run input rate.

In queueing applications we often characterize the system load by the *traffic intensity*, which is the rate in divided by the potential rate out. Here the traffic intensity is

$$\rho \equiv EV_1/\mu .$$

With an infinite-capacity buffer, we need $\rho < 1$ in order for the system to be stable (not blow up in the limit as $t \rightarrow \infty$).

We are able to obtain stochastic-process limits for the workload process by applying the continuous-mapping approach, starting from stochastic-process limits for the centered version of the random walk $\{S_k : k \geq 0\}$. However, to do so when $EX_k \neq 0$, we need to consider a sequence of models indexed by n to achieve the appropriate scaling. In the n^{th} model, we let $X_{n,k}$ be the random-walk step X_k , and we let $EX_{n,k} \rightarrow 0$ as $n \rightarrow \infty$.

There is considerable freedom in the construction of a sequence of models, but from an applied perspective, it suffices to do something simple: We can keep a fixed input process $\{V_k : k \geq 1\}$, but we need to make the output rate μ and the buffer capacity K depend upon n . Let W_k^n denote the workload at the end of period k in model n . Following this plan, for model n the recursion (3.1) becomes

$$W_k^n \equiv \min\{K_n, \max\{0, W_{k-1}^n + V_k - \mu_n\}\}, \quad k \geq 1, \quad (3.2)$$

where K_n and μ_n are the buffer capacity and constant potential one-period output in model n , respectively.

The problem now is to choose the sequences $\{K_n : n \geq 1\}$ and $\{\mu_n : n \geq 1\}$ so that we obtain a nondegenerate limit for an appropriately scaled version of the workload processes $\{W_k^n : k \geq 0\}$. If we choose these sequence of constants appropriately, then the plotter can do the scaling of the workload processes automatically.

Let $S_k^v \equiv V_1 + \cdots + V_k$ for $k \geq 1$ with $S_0^v \equiv 0$. The starting point is a FCLT for the random walk $\{S_k^v : k \geq 0\}$. Suppose that the mean $E[V_k]$ is finite, and let it equal m_v . Then the natural FCLT takes the form

$$\mathbf{S}_n^v \Rightarrow \mathbf{S}^v \quad \text{in } D \quad \text{as } n \rightarrow \infty, \quad (3.3)$$

where

$$\mathbf{S}_n^v(t) \equiv n^{-H}(S_{[nt]}^v - m_v[nt]), \quad 0 \leq t \leq 1, \quad (3.4)$$

the exponent H in the space scaling is a constant satisfying $0 < H < 1$ and $\mathbf{S}^v$ is the limit process. The common case has $H = 1/2$ and $\mathbf{S}^v = \sigma \mathbf{B}$, where

$\mathbf{B}$ is standard Brownian motion. However, as seen for the random walks, if V_k has infinite variance, then we have $1/2 < H < 1$ and the limit process $\mathbf{S}^v$ is a stable Lévy motion (which has discontinuous sample paths). We elaborate on the case with $1/2 < H < 1$ in Section 4.5.

It turns out that a scaled version of the workload process $\{W_k^n : k \geq 0\}$ can be represented directly as the image of a two-sided reflection map applied to a scaled version of the uncentered random walk $\{S_k^n : k \geq 1\}$ with steps $V_k - \mu_n$. In particular,

$$\mathbf{W}_n = \phi_K(\mathbf{S}_n) \quad \text{for all } n \geq 1, \quad (3.5)$$

where

$$\mathbf{W}_n(t) \equiv n^{-H} W_{[nt]}^n, \quad 0 \leq t \leq 1, \quad (3.6)$$

$$\mathbf{S}_n(t) \equiv n^{-H} S_{[nt]}^n, \quad 0 \leq t \leq 1, \quad (3.7)$$

and $\phi_K : D \rightarrow D$ is the *two-sided reflection map*.

In fact, it is a challenge to even define the two-sided reflection map, which we may think of as serving as the continuous-time analog of (3.1) or (3.2); that is done in Sections 5.2 and 14.8; alternatively, see p. 22 of Harrison (1985). Consistent with intuition, it turns out that the two-sided reflection map ϕ_K is continuous on the function space D with appropriate definitions, so that we can apply the continuous-mapping approach with a limit for $\mathbf{S}_n$ in (3.7) to establish the desired limit for $\mathbf{W}_n$. But now we just want to determine how to do the plotting.

The next step is to relate the assumed limit for $\mathbf{S}_n^v$ to the required limit for $\mathbf{S}_n$. For that purpose, note from (3.4) and (3.7) that

$$\mathbf{S}_n(t) = \mathbf{S}_n^v(t) - n^{-H}(\mu_n - m_v)[nt].$$

Hence we have the stochastic-process limit

$$\mathbf{S}_n \Rightarrow \mathbf{S} \quad \text{as } n \rightarrow \infty, \quad (3.8)$$

where

$$\mathbf{S}(t) \equiv \mathbf{S}^v(t) - mt, \quad 0 \leq t \leq 1, \quad (3.9)$$

if and only if

$$n^{-H}(\mu_n - m_v)[nt] \rightarrow mt \quad \text{as } n \rightarrow \infty$$

for each $t > 0$ or, equivalently,

$$(\mu_n - m_v)n^{1-H} \rightarrow m \quad \text{as } n \rightarrow \infty. \quad (3.10)$$

In addition, because of the space scaling by n^H in $\mathbf{S}_n$, we need to let

$$K_n \equiv n^H K . \quad (3.11)$$

Given the scaling in both (3.10) and (3.11), we are able to obtain the FCLT

$$\mathbf{W}_n \Rightarrow \mathbf{W} \equiv \phi_K(\mathbf{S}) , \quad (3.12)$$

where $\mathbf{W}_n$ is given in (3.6), $\mathbf{S}$ is given in (3.9) and ϕ_K is the two-sided reflection map.

The upshot is that we obtain the desired stochastic-process limit for the workload process, and the plotter can automatically do the appropriate scaling, if we let

$$\mu_n \equiv m_v + m/n^{1-H} \quad \text{and} \quad K_n \equiv n^H K \quad (3.13)$$

for any fixed m with $0 \leq m < \infty$ and K with $0 < K \leq \infty$, where H with $0 < H < 1$ is the scaling exponent appearing in (3.4).

At this point, it is appropriate to pause and reflect upon the significance of the scaling in (3.13). First note that time scaling by n (replacing t by nt) and space scaling by n^H (dividing by n^H) is determined by the FCLT in (3.3). Then the output rate and buffer size should satisfy (3.13). Note that the actual buffer capacity K_n in system n must increase, indeed go to infinity, as n increases. Also note that the output rate μ_n approaches m_v as n increases, so that the traffic intensity ρ_n approaches 1 as n increases. Specifically,

$$\rho_n \equiv \frac{E[V_1]}{\mu_n} = \frac{m_v}{m_v + mn^{-(1-H)}} = 1 - (m/m_v)n^{-(1-H)} + o(n^{-(1-H)})$$

as $n \rightarrow \infty$.

The obvious application of the stochastic-process limit in (3.12) is to generate approximations. The direct application of (3.12) is

$$\{n^{-1/\alpha} W_{[nt]}^n : t \geq 0\} \approx \{\mathbf{W}(t) : t \geq 0\} , \quad (3.14)$$

where here $\approx$ means *approximately equal to in distribution*. Equivalently, by unscaling, we obtain the associated approximation (in distribution)

$$\{W_k^n : k \geq 0\} \approx \{n^{1/\alpha} \mathbf{W}(k/n) : k \geq 0\} . \quad (3.15)$$

Approximations such as (3.15), which are obtained directly from stochastic-process limits, may afterwards be refined by making modifications to meet

other criteria, e.g., to match exact expressions known in special cases. Indeed, it is often possible to make refinements that remain asymptotically correct in the heavy-traffic limit, e.g., by including the traffic intensity ρ , which converges to 1 in the limit.

Often the initial goal in support of engineering applications is to develop a suitable approximation. Then heuristic approaches are perfectly acceptable, with convenience and accuracy being the criteria to judge the worth of alternative candidates. Even with such a pragmatic engineering approach, the stochastic-process limits are useful, because they generate initial candidate approximations, often capturing essential features, because the limit often is able to strip away unessential details. Moreover, the limits establish important theoretical reference points, demonstrating asymptotic correctness in certain limiting regimes.

2.3.2. Simulation Examples

Let us now look at two examples.

Example 2.3.1. *Workloads with exponential inputs.*

First let $\{V_k : k \geq 1\}$ be a sequence of IID exponential random variables with mean 1. Then the FCLT in (3.3) holds with $H = 1/2$ and $\mathbf{S}$ being standard Brownian motion $\mathbf{B}$. Thus, from (3.13), the appropriate scaling here is

$$\mu_n \equiv 1 + m/\sqrt{n} \quad \text{and} \quad K_n \equiv \sqrt{n}K. \quad (3.16)$$

To illustrate, we again perform simulations. Due to the recursive definition in (3.2), we can construct and plot the successive workloads just as easily as we constructed and plotted the random walks before. Paralleling our previous plots of random walks, we now plot the first n workloads, using the scaling in (3.16). In Figure 2.1 we plot the first n workloads for the case $H = 1/2$, $m = 1$ and $K = 0.5$ for $n = 10^j$ for $j = 1, \dots, 4$. To supplement Figure 2.1, we show six independent replications for the case $n = 10^4$ in Figure 2.2.

What we see, as n becomes sufficiently large, is standard Brownian motion with drift $-m = -1$ modified by reflecting barriers at 0 and 0.5. Of course, just as for the random-walk plots before, the units on the axes are for the original queueing model. For example, for $n = 10^4$, the buffer capacity is $K_n = 0.5\sqrt{n} = 50$, so that the actual buffer content ranges from 0 to 50, even though the reflected Brownian motion ranges from 0 to 0.5. Similarly, for $n = 10^4$, the traffic intensity is $\rho_n = (1 + n^{-1/2})^{-1} = (1.01)^{-1} \approx 0.9901$ even though the Brownian motion has drift -1 .

Unlike in the previous random-walk plots, the units on the vertical axes in Figure 2.2 are the same for all six plots. That happens because, in all six cases, the workload process takes values ranging from 0 to 50. The upper limit is 50 because for $n = 10^4$ the upper barrier in the queue is $0.5\sqrt{n} = 50$. The clipping at the upper barrier occurs because of overflows.

The traffic intensity 0.99 in Figure 2.2 is admittedly quite high. If we focus instead upon $n = 100$ or $n = 25$, then the traffic intensity is not so extreme, in particular, then $\rho_n = (1 + n^{-1/2})^{-1} = (1.1)^{-1} \approx 0.91$ or $(1.2)^{-1} \approx 0.83$.

In Figures 2.1 and 2.2 we see statistical regularity, just as in the early random-walk plots. Just as in the pairs of figures, (Figures 1.3 and 1.4) and (Figures 1.21 and 1.22), the plots for $n = 10^6$ look just like the plots for $n = 10^4$ when we ignore the units on the axes. The plots show that there should be a stochastic-process limit as $n \rightarrow \infty$. The plots demonstrate that a reflected Brownian motion approximation is appropriate with these parameters.

Moreover, our analysis of the stochastic processes to determine the appropriate scaling shows how we can obtain the stochastic-process limits. Indeed, we obtain the supporting stochastic-process limits for the workload process directly from the established stochastic-process limits for the random walks. In order to make the connection between the random walk and the workload process, we are constrained to use the scaling in (3.16). With that scaling, the plotter directly reveals the statistical regularity. ■

Example 2.3.2. *Workloads with Pareto(3/2) inputs.*

For our second example, we assume that the inputs V_k have a Pareto(p) distribution with finite mean but infinite variance. In particular, we let

$$V_k \equiv U_k^{-1/p} \quad \text{for } p = 3/2, \quad (3.17)$$

just as in case (iii) of (3.5) in Section 1.3.3, which makes the distribution Pareto(p) for $p = 3/2$. Since $H = p^{-1}$ for $p = 3/2$, we need to use different scaling than we did in Example 2.3.1. In particular, instead of (3.16), we now use

$$\mu_n \equiv 1 + m/n^{1/3} \quad \text{and} \quad K_n \equiv n^{2/3}K, \quad (3.18)$$

with $m = 1$ and $K = 0.5$ just as before.

Since the scaling in (3.18) is different from the scaling in (3.16), for any given triple (m, K, n) , the buffer size K_n is now larger, while the output rate differs more from the input rate. Assuming that $m > 0$, the traffic intensity

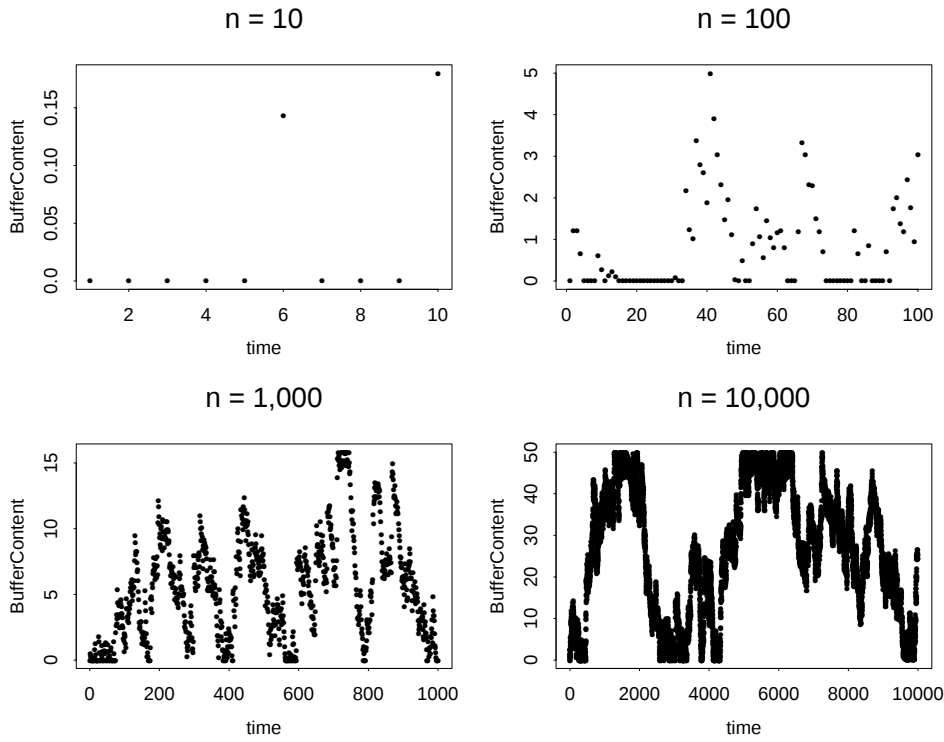


Figure 2.1: Possible realizations of the first n steps of the workload process $\{W_k^n : k \geq 0\}$ with IID exponential inputs having mean 1 for $n = 10^j$ with $j = 1, \dots, 4$. The scaling is as in (3.16) with $m = 1$ and $K = 0.5$.

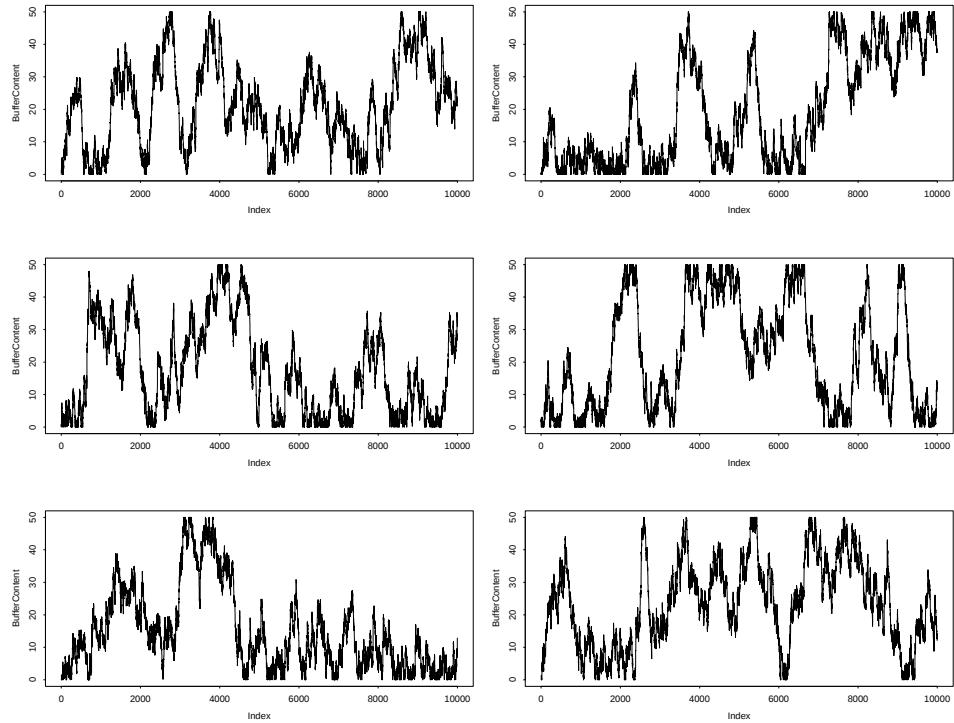


Figure 2.2: Six possible realizations of the first n steps of the workload process $\{W_k^n : k \geq 0\}$ with IID exponential inputs for $n = 10^4$. The scaling is as in (3.16) with $m = 1$ and $K = 0.5$.

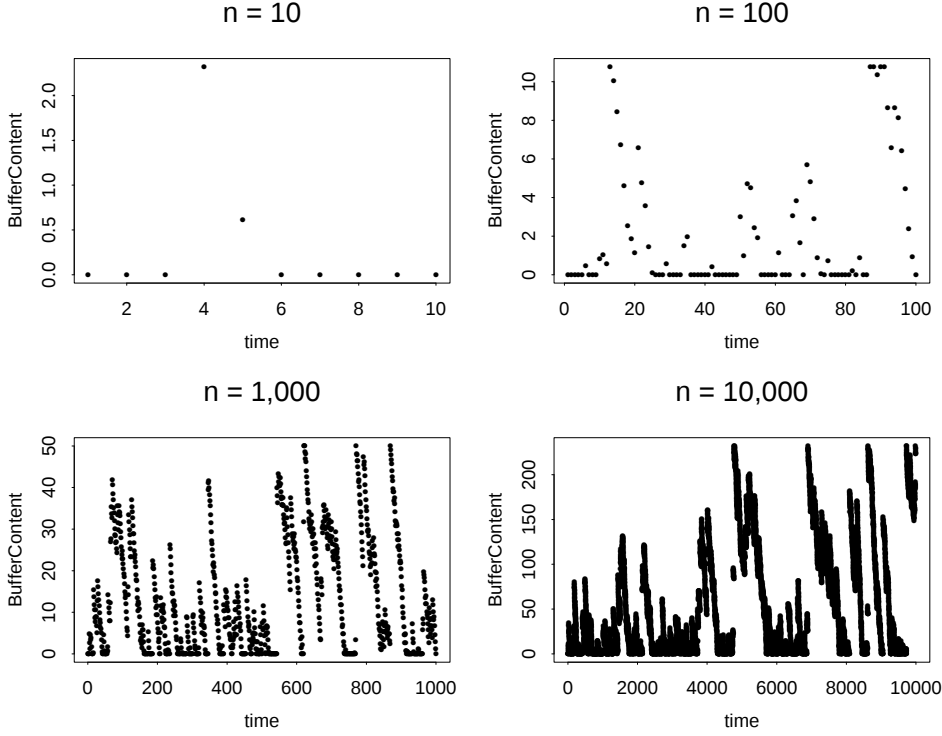


Figure 2.3: Possible realizations of the first n steps of the workload process $\{W_k^n : k \geq 0\}$ with IID $\text{Pareto}(p)$ inputs having $p = 3/2$, mean 3 and infinite variance for $n = 10^j$ with $j = 1, \dots, 4$. The scaling is as in (3.18) with $m = 1$ and $K = 0.5$.

in model n is now lower. That suggests that as H increases the heavy-traffic approximations may perform better at lower traffic intensities.

We plot the first n workloads, using the scaling in (3.18), for $n = 10^j$ for $j = 1, \dots, 4$ in Figure 2.3 for the case $m = 1$ and $K = 0.5$. What we see, as n becomes sufficiently large, is a stable Lévy motion with drift $-m = -1$ modified by reflecting barriers at 0 and 0.5. To supplement Figure 2.3, we show six independent replications for the case $n = 10^4$ in Figure 2.4. As before, the plots for $n = 10^6$ look just like the plots for $n = 10^4$ if we ignore the units on the axes. Just as in Figures 1.20–1.22 for the corresponding random walk, the plots here have jumps. ■

In summary, the workload process $\{W_k\}$ in the queueing model is intimately related to the random walk $\{S_k\}$ with steps being the net inputs

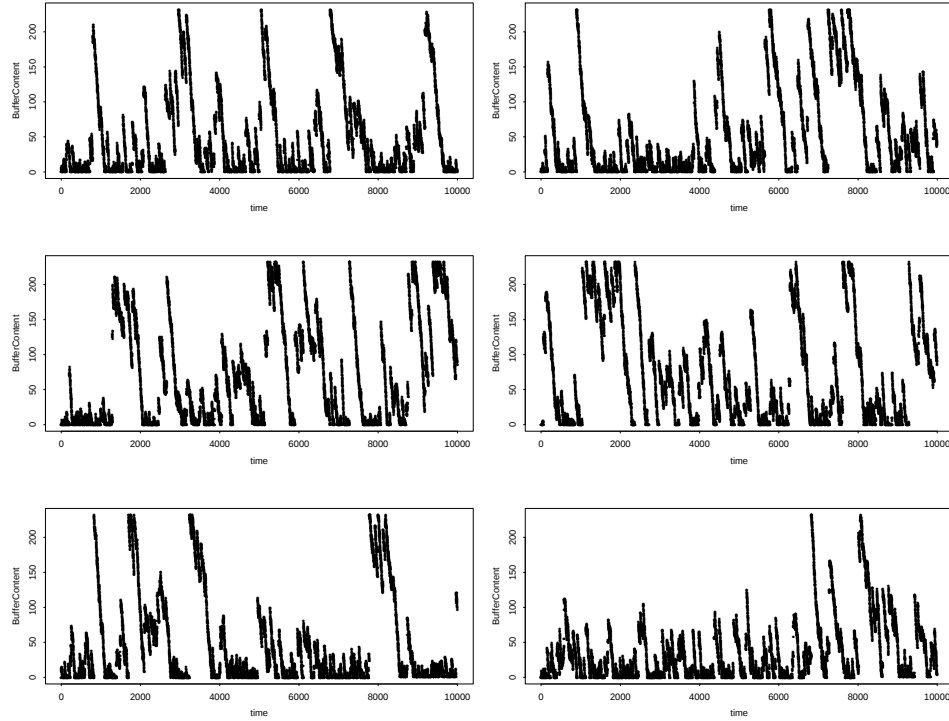


Figure 2.4: Six possible realizations of the first n steps of the workload process $\{W_k^n : k \geq 0\}$ with IID Pareto(p) inputs having $p = 3/2$, mean 3 and infinite variance for $n = 10^4$. The scaling is as in (3.18) with $m = 1$ and $K = 0.5$.

$V_k - \mu$ each period. With appropriate scaling, as in (3.13), which includes the queue being in heavy traffic, stochastic-process limits for a sequence of appropriately scaled workload processes can be obtained directly from associated stochastic-process limits for the underlying random walk.

Moreover, the limit process for the workload process is just the limit process for the random walk modified by having two reflecting barriers. Thus, the workload process in the queue exhibits the same statistical regularity for large sample sizes that we saw for the random walk. Indeed, the random walk is the source of that statistical regularity.

Just as for the random walks, the form of the statistical regularity may lead to the limit process for the workload process having discontinuous sample paths.

2.4. Engineering Significance

In the previous section, we saw that queueing models are closely related to random walks. With the proper (heavy-traffic) scaling, the same forms of statistical regularity that hold for random walks also hold for the workload process in the queueing model. But does it matter? Are there important engineering consequences?

To support an affirmative answer, in this final section we discuss the engineering significance of heavy-traffic stochastic-process limits for queues. First, in Section 2.4.1, we discuss buffer sizing in a switch or router in a communication network. Then, in Section 2.4.2, we discuss scheduling service with multiple sources, as occurs in manufacturing when scheduling production of multiple products on a single machine with setup costs or setup times for switching.

2.4.1. Buffer Sizing

The buffer (waiting space) in a network switch or router tends to be expensive to provide, so that economy dictates it be as small as possible. On the other hand, we want very few lost packets due to buffer overflow.

Queueing models are ideally suited to determine an appropriate buffer size. Let $L(K)$ be the long-run proportion of packets lost as a function of the buffer size K . We might specify a maximum allowable proportion of lost packets, ϵ . Given the function L , we then choose the buffer size K to satisfy the buffer-sizing equation

$$L(K) = \epsilon . \tag{4.1}$$

Classical queueing analysis, using standard models such as in Example 2.3.1, shows that $L(K)$ decays exponentially in K ; specifically, L tends to have an *exponential tail*, satisfying

$$L(K) \sim \alpha e^{-\eta K} \quad \text{as } K \rightarrow \infty \quad (4.2)$$

for asymptotic constants α and η depending upon the model details. (As in (4.6), $\sim$ means asymptotic equivalence. See Remark 5.4.1 for further discussion about asymptotics.)

It is natural to exploit the exponential tail asymptotics for L in (4.2) to generate the approximation

$$L(K) \approx \alpha e^{-\eta K} \quad (4.3)$$

for all K not too small. We then choose K to satisfy the *exponential buffer-sizing equation*

$$\alpha e^{-\eta K} = \epsilon, \quad (4.4)$$

from which we deduce that the target buffer size K^* should be

$$K^* = \eta^{-1} \log(\alpha/\epsilon). \quad (4.5)$$

This analysis shows that the target buffer size should be directly proportional to η^{-1} and $\log \alpha$, and inversely proportional to $\log \epsilon$. It remains to determine appropriate values for the three constants η , α and ϵ , but the general relationships are clear. For example, if $\epsilon = 10^{-j}$, then K^* is proportional to the exponent j , which means that the cost of improving performance (as measured by the increase in buffer size K^* required to make ϵ significantly smaller) tends to be small.

So far, we have yet to exploit heavy-traffic limits. Heavy-traffic limits can play an important role because it actually is difficult to establish the exponential tail asymptotics in (4.2) directly for realistic models. As a first step toward analytic tractability, we may approximate the loss function $L(K)$ by the tail probability $P(W(\infty) > K)$, where $W(\infty)$ is the steady-state workload in the corresponding queue with unlimited waiting space. Experience indicates that the asymptotic form for $L(K)$ tends to be the same as the asymptotic form for the tail probability $P(W(\infty) > K)$ (sometimes with different asymptotic constants). From an applied point of view, we are not too concerned about great accuracy in this step, because the queueing model is crude (e.g., it ignores congestion controls) and the loss proportion $L(K)$ itself is only a rough performance indicator.

As a second step, we approximate $W(\infty)$ in the tail probability $P(W(\infty) > K)$ by the steady-state limit of the approximating process obtained from the heavy-traffic stochastic-process limit. For standard models, the approximating process is reflected Brownian motion, as in Example 2.3.1. Since the steady-state distribution of reflected Brownian motion with one-sided reflection is exponential (see Section 5.7), the heavy-traffic limit provides strong support for the approximations in (4.3)–(4.5) and helps identify approximate values for the asymptotic constants η and α . (The heavy-traffic limits also can generate approximations directly for the loss proportion $L(K)$; e.g., see Section 5.7.) The robustness of heavy-traffic limits (discussed in Chapters 4 and 5) suggests that the analysis should be insensitive to fine system details.

However, the story is not over! Traffic measurements from communication networks present a very different view of the world: These traffic measurements have shown that the traffic carried on these networks is remarkably bursty and complex, exhibiting features such as heavy-tailed probability distributions, strong positive dependence and self-similarity; e.g., see Leland et al. (1994), Garrett and Willinger (1994), Paxson and Floyd (1995), Willinger et al. (1995, 1997), Crovella and Bestavros (1996), Resnick (1997), Adler, Feldman and Taqqu (1998), Barford and Crovella (1998), Crovella, Bestavros and Taqqu (1998), Willinger and Paxson (1998), Park and Willinger (2000), Krishnamurthy and Rexford (2001) and references therein. These traffic studies suggest that different queueing models may be needed.

In particular, the presence of such traffic burstiness can significantly alter the behavior of the queue: *Alternative queueing analysis suggests alternative asymptotic forms for the function L .* Heavy-tailed probability distributions as in Example 2.3.2 lead to a different asymptotic form: When the inputs have power tails, like the Pareto inputs in Example 2.3.2, the function L tends to have a power tail as well: Instead of (4.2), we may have

$$L(K) \sim \alpha K^{-\eta} \quad \text{as } K \rightarrow \infty, \quad (4.6)$$

where again α and η are positive asymptotic constants; see Remark 5.4.1.

The change from the exponential tail in (4.2) to the power tail in (4.6) are contrary to the conclusions made above about the robustness of heavy-traffic approximations. Even though the standard heavy-traffic limits are remarkably robust, there is a limit to the robustness! The traffic burstiness can cause the robustness of the standard heavy-traffic limits to break down. Just as we saw in Example 2.3.2, the burstiness can have a major impact on the workload process.

However, we can still apply heavy-traffic limits: Just as before, we can approximate $L(K)$ by $P(W(\infty) > K)$, where $W(\infty)$ is the steady-state

workload in the corresponding queue with unlimited waiting space. Then we can approximate $W(\infty)$ by the steady-state limit of the approximating process obtained from a heavy-traffic limit. However, when we properly take account of the traffic burstiness, the heavy-traffic limit process is no longer reflected Brownian motion. Instead, as in Example 2.3.2, it may be a reflected stable Lévy motion, for which $P(W(\infty) > K) \sim \alpha K^{-\eta}$. (For further discussion about the power tails, see Sections 4.5, 6.4 and 8.5.) Thus, different heavy-traffic limits support the power-tail asymptotics in (4.6) and yield approximations for the asymptotic constants.

Paralleling (4.3), we can use the approximation

$$L(K) \approx \alpha K^{-\eta} \quad (4.7)$$

for K not too small. Paralleling (4.4), we use the target equation (4.1) and (4.7) to obtain the *power buffer-sizing equation*

$$\alpha K^{-\eta} = \epsilon \quad (4.8)$$

from which we deduce that the *logarithm* of the target buffer size K^* should be

$$\log K^* = \eta^{-1} \log (\alpha/\epsilon) . \quad (4.9)$$

In this power-tail setting, we see that the required buffer size K^* is much more responsive to the parameters η , α and ϵ : Now the logarithm $\log K^*$ is related to the parameters η , α and ϵ the way K^* was before. For example, if $\epsilon = 10^{-j}$, then the logarithm of the target buffer size K^* is proportional to j , which means that the cost of improving performance (as measured by the increase in buffer size K^* required to make ϵ significantly smaller) tends to be large.

And that is not the end! The story is still not over. There are other possibilities: There are different forms of traffic burstiness. In Example 2.3.2 we focused on heavy-tailed distributions for IID inputs, but the traffic measurements also reveal strong dependence. The strong dependence observed in traffic measurements leads to considering fractional-Brownian-motion models of the input, which produce another asymptotic form for the function L ; see Sections 4.6, 7.2 and 8.7. Unlike both the exponential tail in (4.2) and the power tail in (4.5), we may have a *Weibull tail*

$$L(K) \sim \alpha e^{-\eta K^\gamma} \quad \text{as } K \rightarrow \infty \quad (4.10)$$

for positive constants α , η and γ , where $0 < \gamma < 1$; see (8.10) in Section 8.8. The available asymptotic results actually show that

$$P(W(\infty) > K) \sim \alpha K^{-\beta} e^{-\eta K^\gamma} \quad \text{as } K \rightarrow \infty$$

for asymptotic constants η , α and β , where $W(\infty)$ is the steady-state of reflected fractional Brownian motion. Thus, the asymptotic results do not directly establish the asymptotic relation in (4.10), but they suggest the rough approximation

$$L(K) \approx \alpha e^{-\eta K^\gamma} \quad (4.11)$$

for all K not too small and the associated *Weibull buffer-sizing equation*

$$\alpha e^{-\eta K^\gamma} = \epsilon, \quad (4.12)$$

from which we deduce that the γ^{th} power of the target buffer size K^* should be

$$K^{*\gamma} = \eta^{-1} \log(\alpha/\epsilon). \quad (4.13)$$

In (4.13) the γ^{th} power of K^* is related to the parameters α , η and ϵ the way K^* was in (4.5) and $\log K^*$ was in (4.9). Thus, consistent with the intermediate asymptotics in (4.10), since $0 < \gamma < 1$, we have the intermediate buffer requirements in (4.13).

Unfortunately, it is not yet clear which models are most appropriate. Evidence indicates that it depends on the context; e.g., see Heyman and Lakshman (1996, 2000), Ryu and Elwalid (1996), Grossglauser and Bolot (1999), Park and Willinger (2000), Guerin et al. (2000) and Mikosch et al. (2001). Consistent with observations by Sriram and Whitt (1986), long-term variability has relatively little impact on queueing performance when the buffers are small, but can be dramatic when the buffers are large.

Direct traffic measurements are difficult to interpret because they describe the carried traffic, not the offered traffic, and may be strongly influenced by congestion controls such as the Transmission Control Protocol (TCP); see Section 5.2 of Krishnamurthy and Rexford (2001) and Arvidsson and Karlsson (1999). Moreover, the networks and the dominant applications keep changing. For models of TCP, see Padhye et al. (2000), Bu and Towsley (2001), and references therein.

From an engineering perspective, it may be appropriate to ignore congestion controls when developing models for capacity planning. We may wish to provide sufficient capacity so that we usually meet the *offered load* (the original customer demand). When the system is heavily loaded, the controls slow down the stream of packets. From a careful analysis of traffic measurements, we may be able to reconstruct the intended flow. (For further discussion about offered-load models, see Remark 10.3.1.) However, heavy-traffic limits can also describe the performance with congestion-controlled sources, as shown by Das and Srikant (2000).

Our goal in this discussion, and more generally in the book, is not to draw engineering conclusions, but to describe an approach to engineering problems: Heavy-traffic limits yield simple approximations that can be used in engineering applications involving queues. Moreover, nonstandard heavy-traffic limits can capture the nonstandard features observed in network traffic. The simple analysis above shows that the consequences of the model choice can be dramatic, making order-of-magnitude differences in the predicted buffer requirements.

When the analysis indicates that very large buffers are required, instead of actually providing very large buffers, we may conclude that buffers are relatively ineffective for improving performance. Instead of providing very large buffers, we may choose to increase the available bandwidth (processing rate), introduce scheduling to reduce the impact of heavy users upon others, or regulate the source inputs (see Example 9.8.1). Indeed, all of these approaches are commonly used in practice. It is common to share the bandwidth among sources using a “fair queueing” discipline. Fair queueing disciplines are variants of the head-of-line processor-sharing discipline, which gives each of several active sources a guaranteed share of the available bandwidth. See Demers, Keshav and Shenker (1989), Greenberg and Madras (1992), Parekh and Gallager (1993, 1994), Anantharam (1999) and Borst, Boxma and Jelenković (2000).

Many other issues remain to be considered: First, given any particular asymptotic form, it remains to estimate the asymptotic constants. Second, it remains to determine how the queueing system scales with increasing load. Third, it may be more appropriate to consider the transient or time-dependent performance measures instead of the customary steady-state performance measures. Fourth, it may be necessary to consider more than a single queue in order to capture network effects. Finally, it may be necessary to create appropriate controls, e.g., for scheduling and routing. Fortunately, for all these problems, and others, heavy-traffic stochastic-process limits can come to our aid.

2.4.2. Scheduling Service for Multiple Sources

In this final subsection we discuss *the engineering significance of the time-and-space scaling* that occurs in heavy-traffic limits for queues. The heavy-traffic scaling was already discussed in Section 2.3; now we want to point out its importance for system control.

We start by extending the queueing model in Section 2.3: Now we assume that there are inputs each time period from m separate sources. We let each

source have its own infinite-capacity buffer, and assume that the work in each buffer is served in order of arrival, but otherwise we leave open the order of service provided to the different sources. As before, we can think of there being a single server, but now the server has to switch from queue to queue in order to perform the service, with there being a setup cost or a setup time to do the switching.

We initially assume that the server can switch from queue to queue instantaneously (within each discrete time period), but we assume that there are switchover costs for switching. To provide motivation for switching, we also assume that there are source-dependent holding costs for the workloads. To specify a concrete optimization problem, let W_k^i denote the source- i workload in its buffer at the end of period k and let $S_k^{i,j}$ be the number of switches from queue i to queue j in the first k periods. Let the total cost incurred in the first k periods be the sum of the total holding cost and the total switching cost, i.e.,

$$C_k \equiv H_k + S_k ,$$

where

$$H_k \equiv \sum_{i=1}^m \sum_{j=1}^k h_i W_j^i$$

and

$$S_k \equiv \sum_{i=1}^m \sum_{j=1}^m c_{i,j} S_k^{i,j} ,$$

where h_i is the source- i holding cost per period and $c_{i,j}$ is the switching cost per switch from source i to source j . Our goal then may be to choose a switching policy that minimizes the long-run average expected cost

$$\bar{C} \equiv \lim_{k \rightarrow \infty} k^{-1} E[C_k] .$$

This is a difficult control problem, even under the regularity condition that the inputs come from m independent sequences of IID random variables with finite means m_v^i . Under that regularity condition, the problem can be formulated as a *Markov sequential decision process*; e.g., see Puterman (1994): The state at the beginning of period $k + 1$ is the workload vector $(W_k^1, \dots, W_k^m)$ and the location of the server at the end of period k . An action is a specification of the sequence of queues visited and the allocation of the available processing per period, μ , during those visits. Both the state and action spaces are uncountably infinite, but we could make reasonable simplifying assumptions to make them finite.

To learn how we might approach the optimization problem, it is helpful to consider a simple scheduling policy: A *polling* policy serves the queues to exhaustion in a fixed cyclic order, with the server starting each period where it stopped the period before. We assume that the server keeps working until either its per-period capacity μ is exhausted or all the queues are empty.

There is a large literature on polling models; see Takagi (1986) and Boxma and Takagi (1992). For classical polling models, there are analytical solutions, which can be solved numerically. For those models, numerical transform inversion is remarkably effective; see Choudhury and Whitt (1996). However, analytical tractability is soon lost as model complexity increases, so there is a need for approximations.

The polling policy is said to be a *work-conserving service policy*, because the server continues serving as long as there is work in the system yet to be done (and service capacity yet to provide). An elementary, but important, observation is that the total workload process for any work-conserving policy is identical to the workload process with a single shared infinite-capacity buffer. Consequently, the heavy-traffic limit described in Section 2.3 in the special case of an infinite buffer ($K = \infty$) also holds for the total-workload process with polling; i.e., with the FCLT for the cumulative inputs in (3.3) and the heavy-traffic scaling in (3.10), we have the heavy-traffic limit for the scaled total-workload processes in (3.12), with the two-sided reflection map ϕ_K replaced by the one-sided reflection map. Given the space scaling by n^H and the time scaling by n , where $0 < H < 1$, the unscaled total workload at any time in the n^{th} system is of order n^H and changes significantly over time intervals having length of order n .

The key observation is that the time scales are very different for the individual workloads at the source buffers. First, the individual workloads are bounded above by the total workload. Hence the unscaled individual workloads are also of order n^H . Clearly, the mean inputs must satisfy the relation

$$m_v = m_{v,1} + \cdots + m_{v,m} .$$

Assuming that $0 < m_{v,i} < m_v$ for all i , we see that *each source by itself is not in heavy traffic when the server is dedicated to it*: With the heavy-traffic scaling in (3.10), the total traffic intensity approaches 1, i.e.,

$$\rho_n \equiv m_v / \mu_n \uparrow 1 \quad \text{as } n \rightarrow \infty ,$$

but the instantaneous traffic intensity for source i when the server is devoted to it converges to a limit less than 1, i.e.,

$$\rho_{n,i} \equiv m_{v,i} / \mu_n \uparrow m_{v,i} / m_v \equiv \rho_i^* < 1 .$$

Since each source alone is not in heavy-traffic when the server is working on that source, the net output is at a constant positive rate when service is being provided, even in the heavy-traffic limit. Thus the server processes the order n^H unscaled work there in order n^H time, by the law of large numbers (see Section 5.3).

The upshot is that the unscaled individual workloads change significantly in order n^H time whenever the server is devoted to them, and the server cycles through the m queues in order n^H time, whereas the unscaled total workload changes significantly in order n time. Since $H < 1$, in the heavy-traffic limit the individual workloads change on a faster time scale. Thus, in the heavy-traffic limit we obtain a *separation of time scales*: When we consider the evolution of the individual workload processes in a short time scale, we can act as if the total workload is fixed.

Remark 2.4.1. *The classic setting: NCD Markov chains.* The separation of time scales in the polling model is somewhat surprising, because it occurs in the heavy-traffic limit. In other settings, a separation of time scales is more evident. With computers and communication networks, the relevant time scale for users is typically seconds, while the relevant time scale for system transactions is typically milliseconds. For those systems, engineers know that time scales are important.

There is a long tradition of treating different time scales in stochastic models using nearly-completely-decomposable (NCD) Markov chains; see Courtois (1977). With a NCD Markov chain, the state space can be decomposed into subsets such that most of the transitions occur between states in the same subset, and only rarely does the chain move from one subset to another. In a long time scale, the chain tends to move from one local steady-state regime to another, so that the long-run steady-state distribution is an appropriate average of the local steady-state distributions.

However, different behavior can occur if the chain does not approach steady-state locally within a subset. For example, that occurs in an infinite-capacity queue in a slowly changing environment when the queue is unstable in some environment states. Heavy-traffic limits for such queues were established by Choudhury, Mandelbaum, Reiman and Whitt (1997). Even though the queue content may ultimately approach a unique steady-state distribution, the local instability may cause significant fluctuations in an intermediate time scale. The transient behavior of the heavy-traffic limit process captures this behavior over the intermediate time scale. ■

For the polling model, the separation of time scales suggests that in the heavy-traffic limit, given the fixed scaled total workload $\mathbf{W}_n(t) = w$,

in the neighborhood of time t the vector of scaled individual workloads $(\mathbf{W}_n^1(t), \dots, \mathbf{W}_n^m(t))$ rapidly traverses a deterministic piecewise-linear trajectory through points $(w^1, \dots, w^m)$ in the hyperplane in $\mathbb{R}^m$ with $w^1 + \dots + w^m = w$. For example, with three identical sources served in numerical cyclic order, the path is piecewise-linear, passing through the vertices $(2w/3, w/3, 0)$, $(0, 2w/3, w/3)$ and $(w/3, 0, 2w/3)$, corresponding to the instants the server is about to start service on sources 1, 2 and 3, respectively. In general, identifying the vertices is somewhat complicated, but the experience of each source is clear: it builds up to its peak workload at constant rate and then returns to emptiness at constant rate. And it does this many times before the total workload changes significantly. Hence at any given time its level can be regarded as uniformly distributed over its range.

As a consequence, we anticipate a *heavy-traffic averaging principle*: We should have a limit for the average of functions of the scaled individual workloads; i.e., for any $s, h > 0$ and any continuous real-valued function f ,

$$h^{-1} \int_s^{s+h} f(\mathbf{W}_n^i(t)) dt \Rightarrow h^{-1} \int_s^{s+h} \left(\int_0^1 f(a_i u \mathbf{W}(t)) du \right) dt, \quad (4.14)$$

where a_i is a constant satisfying $0 < a_i \leq 1$ for $1 \leq i \leq m$. In words, the time-average of the scaled individual-source workload process over the time interval $[s, s+h]$ approaches the corresponding time-average of a proportional space-average of the limit $\mathbf{W}$ for the scaled total workload process. (For other instances of the averaging principle, see Anisimov (1993) and Freidlin and Wentzell (1993).)

This heavy-traffic averaging principle was rigorously established for the case of two queues by Coffman, Puhalskii and Reiman (1995) for a slightly different model in the Brownian case, with $H = 1/2$ and $\mathbf{W}$ reflected Brownian motion. They also determined the space-scaling constants a_i appearing in (4.14) for m sources: They showed that

$$a_i = \frac{\rho_i^* (1 - \rho_i^*)}{\sum_{1 \leq j < k \leq m} \rho_j^* \rho_k^*}, \quad (4.15)$$

where ρ_i^* is the limiting source- i traffic intensity, i.e., $\rho_i^* \equiv m_{v,i}/m_v$ for our model. The upper limits a_i depend only on the means $m_{v,j}$, $1 \leq j \leq m$. For $m = 2$, $a_i = 1$; for m identical sources, $a_i = 2/m$. The variability affects the limit in (4.14) only through the scaling and the one-dimensional limit process $\mathbf{W}$.

Coffman, Puhalskii and Reiman (1998) also considered the two-queue polling model with unscaled switchover times. Even though the switchover

times are asymptotically negligible in the heavy-traffic scaling, they have a significant impact because the relative amount of switching increases as the total workload decreases. Coffman, Puhalskii and Reiman (1998) show that the heavy-traffic averaging principle is still valid with switchover times, with the scaled total workload processes converging to a Bessel diffusion process, which has state-dependent drift of the form $-a + b/x$ for positive constants a and b . (For additional heavy-traffic limits for polling models, see van der Mei and Levy (1997) and van der Mei (2000).)

Even though the polling models have yet to be analyzed for nonstandard scaling, with $H \neq 1/2$ and $\mathbf{W}$ not a diffusion process, it is evident that the heavy-traffic averaging principle still applies. We can anticipate that the other forms of variability (associated with heavy tails and strong dependence) affect the heavy-traffic limit only through the limit process $\mathbf{W}$.

The separation of time scales provides a way to attack complicated service control problems such as the one formulated at the beginning of this subsection. Even if all the desired supporting mathematics cannot be established, the heavy-traffic limits provide a useful perspective for approximately solving these problems. The heavy-traffic averaging principle reduces the dimension of the state-space in the control problem. It provides a form of *state-space collapse*; see Reiman (1984b), Harrison and van Mieghem (1997), Bramson (1998) and Williams (1998b). It lets us focus on the single process that is the heavy-traffic limit for the scaled total-workload process. For natural classes of service policies, we can express the local cost rate associated with a fixed total workload and then determine an expression for the long-run average total cost as a function of the controls that produces a tractable optimization problem. In the more challenging cases it may be necessary to apply numerical methods to solve the optimization problem, as in Kushner and Dupuis (2000).

By now, there has been substantial work on this heavy-traffic approach to scheduling, yielding excellent results. We do not try to tell the story here; instead we refer to Reiman and Wein (1998), Markowitz, Reiman and Wein (2000), Markowitz and Wein (2001) and Kushner (2001).

For these more complicated control problems, there are many open technical problems: It remains to establish the heavy-traffic averaging principle in more complicated settings and it remains to show that the derived policies are indeed asymptotically optimal in the heavy-traffic limit. Markowitz et al. (2000, 2001) restrict attention to dynamic cyclic policies in which each source is served once per cycle in the same fixed order. It is easy to construct examples in which larger classes of policies are needed: With three sources, it may be necessary to serve one source more frequently; e.g., the

cycle $(1, 2, 1, 3)$ may be much better than either $(1, 2, 3)$ or $(1, 3, 2)$.

Nevertheless, the practical value of the heavy-traffic approach is well established: Numerical comparisons have shown that the policies generated from the heuristic heavy-traffic analysis perform well for systems under normal loading. Moreover, the heavy-traffic analysis produces important insight about the control problem, as illustrated by concluding remarks on p. 268 of Markowitz and Wein (2001) about the way model features – setups, due dates and product mix – affect the structure of policies. And there is opportunity for further work along these lines.

Heavy-traffic analysis has also been applied to other queueing control problems. We have discussed the scheduling of service for multiple sources by a single server. We may instead have to schedule and route input from multiple sources to several possible servers; see Bell and Williams (2001), Harrison and Lopez (1999) and references therein. More generally, we may have multiclass processing networks; see Harrison (1988, 2000, 2001a,b), Kumar (2000) and references therein.

In conclusion, the successful application of heavy-traffic analysis to these classic operations-research stochastic scheduling problems provides ample evidence that heavy-traffic stochastic-process limits for queues have engineering significance.

Chapter 5

Heavy-Traffic Limits for Fluid Queues

5.1. Introduction

In this chapter we see how the continuous-mapping approach can be applied to establish heavy-traffic stochastic-process limits for queueing models, and how those heavy-traffic stochastic-process limits, in turn, can be applied to obtain approximations for queueing processes and gain insight into queueing performance.

To establish the heavy-traffic stochastic-process limits, the general idea is to represent the queueing “content” process of interest as a reflection of a corresponding net-input process. For single queues with unlimited storage capacity, a one-sided one-dimensional reflection map is used; for single queues with finite storage capacity, a two-sided one-dimensional reflection map is used. These one-dimensional reflection maps are continuous as maps from D to D with all the principal topologies considered by virtue of results in Sections 13.5 and 14.8. Hence, FCLT’s for scaled net-input processes translate into corresponding FCLT’s for scaled queueing processes.

Thus we see that the relatively tractable heavy-traffic approximations can be regarded as further instances of the statistical regularity stemming from the FCLT’s in Chapter 4. The FCLT for the scaled net-input processes may be based on Donsker’s theorem in Section 4.3 and involve convergence to Brownian motion; then the limit process for the scaled queueing processes is reflected Brownian motion (RBM). Alternatively, the FCLT for the scaled net-input processes may be based on one of the other FCLT’s in Sections

4.5 – 4.7 and involve convergence to a different limit process; then the limit process for the scaled queueing processes is the reflected version of that other limit process.

For example, when the net-input process can be constructed from partial sums of IID random variables with heavy-tailed distributions, Section 4.5 implies that the scaled net-input processes converge to a stable Lévy motion; then the limit process for the queueing processes is a reflected stable Lévy motion. The reflected stable Lévy motion heavy-traffic limit describes the effect of the extra burstiness due to the heavy-tailed distributions.

As indicated in Section 4.6, it is also possible to have more burstiness due to strong positive dependence or less burstiness due to strong negative dependence. When the net-input process has such strong dependence with light-tailed distributions, the scaled net-input processes may converge to fractional Brownian motion; then the limit process for the scaled queueing processes is reflected fractional Brownian motion.

In this chapter, attention will be focused on the “classical” Brownian approximation involving RBM and its application. For example, in Section 5.8 we show how the heavy-traffic stochastic-process limit with convergence to RBM can be used to help plan queueing simulations, i.e., to estimate the required run length to achieve desired statistical precision, as a function of model parameters. Reflected stable Lévy motion will be discussed in Sections 8.5 and 9.7, while reflected fractional Brownian motion will be discussed in Sections 8.7 and 8.8.

In simple cases, the continuous-mapping approach applies directly. In other cases, the required argument is somewhat more complicated. A specific simple case is the discrete-time queueing model in Section 2.3. In that case, the continuous-mapping argument applies directly: FCLT’s for the partial sums of inputs V_k translate immediately into associated FCLT’s for the workload (or buffer-content) process $\{W_k\}$, exploiting the continuity of the two-sided reflection map. The continuous-mapping approach applies directly because, as indicated in (3.5) in Chapter 1, the scaled workload process is exactly the reflection of the scaled net-input process, which itself is a scaled partial-sum process. Thus all the stochastic-process limits in Chapter 4 translate into corresponding heavy-traffic stochastic-process limits for the workload process in Section 2.3.

In this chapter we see how the continuous-mapping approach works with related continuous-time fluid-queue models. We start considering fluid queues, instead of standard queues (which we consider in Chapter 9), because fluid queues are easier to analyze and because fluid queues tend to serve as initial “rough-cut” models for a large class of queueing systems.

The fluid-queue models have recently become popular because of applications to communication networks, but they have a long history. In the earlier literature they are usually called dams or stochastic storage models; see Moran (1959) and Prabhu (1998). In addition to queues, they have application to inventory and risk phenomena.

In this chapter we give proofs for the theorems, but the emphasis is on the statement and applied significance of the theorems. The proofs illustrate the continuous-mapping approach for establishing stochastic-process limits, exploiting the useful functions introduced in Section 3.5. Since the proofs draw on material from later chapters, upon first reading it should suffice to focus, first, on the theorem statements and their applied significance and, second, on the general flow of the argument in the proofs.

5.2. A General Fluid-Queue Model

In a fluid-queue model, a divisible commodity (fluid) arrives at a storage facility where it is stored in a buffer and gradually released. We consider an *open model* in which fluid arrives *exogenously* (from outside). For such open fluid-queue models, we describe the buffer content over time. In contrast, in a standard queueing model, which we consider in Chapter 9, individual customers (or jobs) arrive at a service facility, possibly wait, then receive service and depart. For such models, we count the number of customers in the system and describe the experience of individual customers. The fluid queue model can be used to represent the unfinished work in a standard queueing model. Then the input consists of the customer service requirements at their arrival epochs. And the unfinished work declines at unit rate as service is provided.

In considering fluid-queue models, we are motivated to a large extent by the need to analyze the performance of evolving communication networks. Since data carried by these networks are packaged in many small packets, it is natural to model the flow as fluid, i.e., to think of the flow coming continuously over time at a random rate. A congestion point in the network such as a switch or router can be regarded as a queue (dam or stochastic storage model), where input is processed at constant or variable rate (the available bandwidth). Thus, we are motivated to consider fluid queues. However, we should point out that other approaches besides queueing analysis are often required to engineer communication networks; to gain perspective, see Feldmann et al. (2000, 2001) and Krishnamurthy and Rexford (2001).

5.2.1. Input and Available-Processing Processes

In this section we consider a very general model: We consider a single fluid queue with general input and available-processing (or service) processes. For any $t > 0$, let $C(t)$ be the cumulative input of fluid over the interval $[0, t]$ and let $S(t)$ be the cumulative available processing over the interval $[0, t]$. If there is always fluid to process during the interval $[0, t]$, then the quantity processed during $[0, t]$ is $S(t)$. We assume that $\{C(t) : t \geq 0\}$ and $\{S(t) : t \geq 0\}$ are real-valued stochastic processes with nondecreasing nonnegative right-continuous sample paths. But at this point we make no further structural or stochastic assumptions.

A common case is processing at a constant rate μ whenever there is fluid to process; then

$$S(t) = \mu t, \quad t \geq 0. \quad (2.1)$$

More generally, we could have input and output at random rates. Then

$$C(t) = \int_0^t R_i(s) ds \quad \text{and} \quad S(t) = \int_0^t R_o(s) ds, \quad t \geq 0, \quad (2.2)$$

where $\{R_i(t) : t \geq 0\}$ and $\{R_o(t) : t \geq 0\}$ are nonnegative real-valued stochastic processes with sample paths in D . For example, it is natural to have maximum possible input and processing rates ν_i and ν_o . Then, in addition to (2.2), we would assume that

$$0 \leq R_i(t) \leq \nu_i \quad \text{and} \quad 0 \leq R_o(t) \leq \nu_o \quad \text{for all } t \text{ w.p.1.} \quad (2.3)$$

With (2.2), the stochastic processes C and S have continuous sample paths. We regard that as the standard case, but we allow C and S to be more general.

With the general framework, the discrete-time fluid-queue model in Section 2.3 is actually a special case of the continuous-time fluid-queue model considered here. The previous discrete-time fluid queue is put in the present framework by letting

$$C(t) \equiv \sum_{k=1}^{\lfloor t \rfloor} V_k \quad \text{and} \quad S(t) \equiv \mu \lfloor t \rfloor, \quad t \geq 0,$$

where $\lfloor t \rfloor$ is the greatest integer less than or equal to t .

5.2.2. Infinite Capacity

We will consider both the case of unlimited storage space and the case of finite storage space. First suppose that there is unlimited storage space. Let $W(t)$ represent the *workload* (or buffer content, i.e., the quantity of fluid waiting to be processed) at time t . Note that we can have significant fluid flow without ever having any workload. For example, if $W(0) = 0$, $C(t) = \lambda t$ and $S(t) = \mu t$ for all $t \geq 0$, where $\lambda < \mu$, then fluid is processed continuously at rate λ , but $W(t) = 0$ for all t . However, if C is a pure-jump process, then the processing occurs only when $W(t) > 0$. (The workload or virtual-waiting-time process in a standard queue is a pure-jump process.)

The workload $W(t)$ can be defined in terms of an *initial workload* $W(0)$ and a *net-input process* $C(t) - S(t)$, $t \geq 0$, via a *potential-workload process*

$$X(t) \equiv W(0) + C(t) - S(t), \quad t \geq 0, \quad (2.4)$$

by applying the *one-dimensional reflection map* to X , i.e., by letting

$$W(t) \equiv \phi(X)(t) \equiv X(t) - \inf_{0 \leq s \leq t} \{X(s) \wedge 0\}, \quad t \geq 0, \quad (2.5)$$

where $a \wedge b = \min\{a, b\}$.

We could incorporate the initial workload $W(0)$ into the cumulative-input process $\{C(t) : t \geq 0\}$ by letting $C(0) = W(0)$. Then X would simply be the net-input process. However, we elect not to do this, because it is convenient to treat the initial conditions separately in the limit theorems.

The potential workload represents what the workload would be if we ignored the emptiness condition, and assumed that there is always output according to the available-processing process S . Then the workload at time t would be $X(t)$: the sum of the initial workload $W(0)$ plus the cumulative input $C(t)$ minus the cumulative output $S(t)$. Since emptiness may sometimes prevent output, we have definition (2.5).

Formula (2.5) is easy to understand by looking at a plot of the potential workload process $\{X(t) : t \geq 0\}$, as shown in Figure 5.1. Figure 5.1 shows a possible sample path of X when $S(t) = \mu t$ for $t \geq 0$ w.p.1 and there is only one on-off source that alternates between busy periods and idle periods, having input rate $r > \mu$ during busy periods and rate 0 during idle periods. Hence the queue alternates between net-input rates $r - \mu > 0$ and $-\mu < 0$. The plot of the potential workload process $\{X(t) : t \geq 0\}$ also can be interpreted as a plot of the actual workload process if we redefine what is meant by the origin. For the workload process, the origin is either 0, if X has not become negative, or the lowest point reached by X . The position of the origin for W is shown by the shaded dashed line in Figure 5.1.

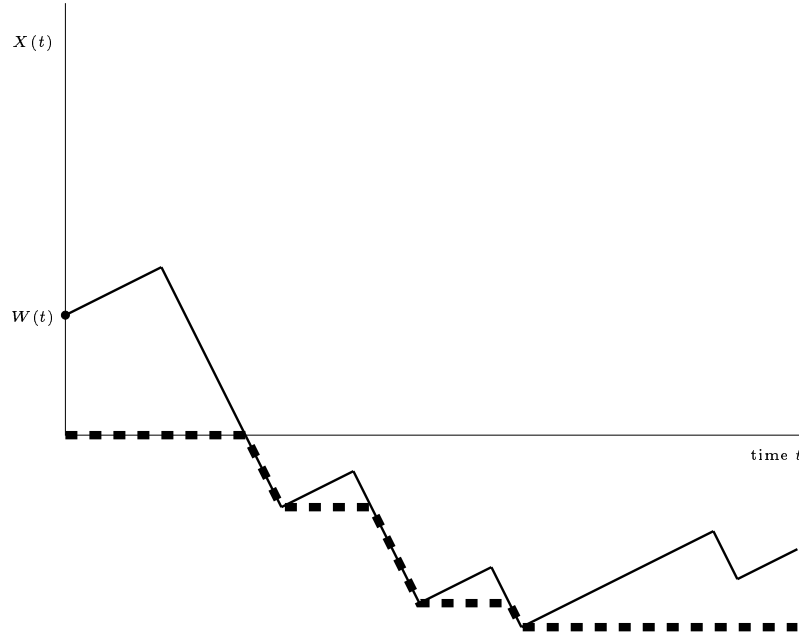


Figure 5.1: A possible realization of the potential workload process $\{X(t) : t \geq 0\}$ and the actual workload process $\{W(t) : t \geq 0\}$ with unlimited storage capacity: The actual workload process appears if the origin is the heavy shaded dashed line; i.e., solid line - dashed line = actual workload.

An important observation is that the single value $W(t)$, for any $t > 0$, depends on the initial segment $\{X(s) : 0 \leq s \leq t\}$. To know $W(t)$, it is not enough to know the single value $X(t)$. However, by (2.5) it is evident that, for any $t > 0$, both $W(t)$ and the initial segment $\{W(s) : 0 \leq s \leq t\}$ are functions of the initial segment $\{X(s) : 0 \leq s \leq t\}$. With appropriate definitions, the reflection map in (2.5) taking the modified net-input process $\{X(t) : t \geq 0\}$ into the workload processes $\{W(t) : t \geq 0\}$ is a continuous function on the space of sample paths; see Section 13.5. Thus, by exploiting the continuous mapping theorem in a function space setting, a limit for a sequence of potential workload processes will translate into a corresponding limit for the associated sequence of workload processes.

Remark 5.2.1. *Model generality.* It may be hard to judge whether the fluid queue model we have introduced is exceptionally general or restrictive. It depends on the perspective: On the one hand, the model is very general

because the basic stochastic processes C and S can be almost anything. We illustrate in Chapter 8 by allowing the input C to come from several on-off sources. We are able to treat that more complex model as a special case of the model studied here. On the other hand, the model is also quite restrictive because we assume that the workload stochastic process is directly a reflection of the potential-workload stochastic process. That makes the continuous-mapping approach especially easy to apply. In contrast, as we will see in Chapter 9, it is more difficult to treat the queue-length process in the standard single-server queue without special Markov assumptions. However, additional mathematical analysis shows that the model discrepancy is asymptotically negligible: In the heavy-traffic limit, the queue-length process in the standard single-server queue behaves as if it could be represented directly as a reflection of the associated net-input process. And similar stories hold for other models. The fluid model here is attractive, not only because it is easy to analyze, but also because it captures the essential nature of more complicated models. ■

The general goal in studying this fluid-queue model is to understand how assumed behavior of the basic stochastic processes C and S affects the workload stochastic process W . For example, assuming that the net-input process $C - S$ has stationary increments and negative drift, under minor regularity conditions (see Chapter 1 of Borovkov (1976)), the workload $W(t)$ will have a limiting steady-state distribution. We want to understand how that steady-state distribution depends on the stochastic processes C and S . We also want to describe the transient (time-dependent) behavior of the workload process. Heavy-traffic limits can produce robust approximations that may be useful even when the queue is not in heavy traffic.

We now want to consider the case of a finite storage capacity, but before defining the finite-capacity workload process, we note that the one-sided reflection map in (2.5) can be expressed in an alternative way, which is convenient for treating generalizations such as the finite-capacity model and fluid networks; see Chapter 14 and Harrison (1985) for more discussion. Instead of (2.5), we can write

$$W(t) \equiv \phi(X)(t) \equiv X(t) + L(t), \quad (2.6)$$

where X is the potential workload process in (2.4) and $\{L(t) : t \geq 0\}$ is a nondecreasing “regulator” process that increases only when $W(t) = 0$, i.e., such that

$$\int_0^t W(s) dL(s) = 0, \quad t \geq 0. \quad (2.7)$$

From (2.5), we know that

$$L(t) = - \inf_{0 \leq s \leq t} \{X(s) \wedge 0\}, \quad t \geq 0. \quad (2.8)$$

It can be shown that the characterization of the reflection map via (2.6) and (2.7) is equivalent to (2.5). For a detailed proof and further discussion, see Chapter 14, which focuses on the more complicated multidimensional generalization.

5.2.3. Finite Capacity

We now modify the definition in (2.6) and (2.7) to construct the finite-capacity workload process. Let the buffer capacity be K . Now we assume that any input that would make the workload process exceed K is lost. Let

$$W(t) \equiv \phi_K(X)(t) \equiv X(t) + L(t) - U(t), \quad t \geq 0, \quad (2.9)$$

where again $X(t)$ is the potential workload process in (2.4), the initial condition is now assumed to satisfy $0 \leq W(0) \leq K$, and $L(t)$ and $U(t)$ are both nondecreasing processes. The *lower-boundary regulator process* $L \equiv \psi_L(X)$ increases only when $W(t) = 0$, while the *upper-boundary regulator process* $U \equiv \psi_U(X)$ increases only when $W(t) = K$; i.e., we require that

$$\int_0^t W(s) dL(s) = \int_0^t [K - W(s)] dU(s) = 0, \quad t \geq 0. \quad (2.10)$$

The random variable $U(t)$ represents the quantity of fluid lost (the overflow) during the interval $[0, t]$. We are often interested in the overflow process $\{U(t) : t \geq 0\}$ as well as the workload process $\{W(t) : t \geq 0\}$.

Note that we can regard the infinite-capacity model as a special case of the finite-capacity model. When $K = \infty$, we can regard the second integral in (2.10) as implying that $U(t) = 0$ for all $t \geq 0$.

Closely paralleling Figure 5.1, for the finite-capacity model we can also depict possible realizations of the processes X and W together, as shown in Figure 5.2. As before, the potential workload process is plotted directly, but we also see the workload (buffer content) process W if we let the origin and upper barrier move according to the two heavily shaded dashed lines, which remain a distance K apart. Decreases in the dashed lines correspond to increases in the lower-barrier regulator process L , while increases in the shaded lines correspond to increases in the upper-barrier regulator process U . From the Figure 5.2, the validity of (2.9) and (2.10) is evident. Furthermore, it

is evident that the two-sided reflection in (2.9) can be defined by successive applications of the one-sided reflection map in (2.5) and (2.6) corresponding to the lower and upper barriers separately. For further discussion, see Section 14.8.

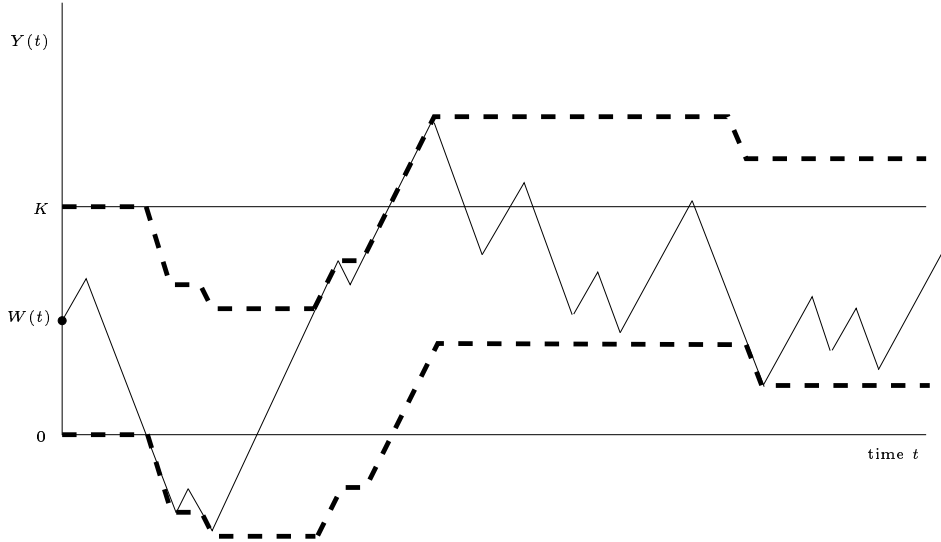


Figure 5.2: A possible realization of the potential workload process $\{X(t) : t \geq 0\}$ and the actual workload process $\{W(t) : t \geq 0\}$ with finite storage capacity K : The actual workload process appears if the origin and upper limit are the heavily shaded dashed lines always a distance K apart. As in Figure 5.1, solid line - lower dashed line = actual workload.

As in the infinite-capacity case, given K , the initial segment $\{W(s), L(s), U(s) : 0 \leq s \leq t\}$ depends on the potential-workload process X via the corresponding initial segment $\{X(s) : 0 \leq s \leq t\}$. Again, under regularity conditions, the reflection map in (2.9) taking $\{X(t) : t \geq 0\}$ into $\{(W(t), L(t), U(t)) : t \geq 0\}$ is a continuous function on the space of sample paths (mapping initial segments into initial segments). Thus, stochastic-process limits for X translate into stochastic-process limits for (W, L, U) , by exploiting the continuous-mapping approach with the full reflection map (ϕ_K, ψ_L, ψ_U) in a function space setting.

Let $D(t)$ represent the amount of fluid processed (not counting any overflow) during the time interval $[0, t]$. We call $\{D(t) : t \geq 0\}$ the *departure process*.

From (2.4) and (2.9),

$$\begin{aligned} D(t) &= W(0) + C(t) - W(t) - U(t) \\ &= S(t) - L(t), \quad t \geq 0. \end{aligned} \quad (2.11)$$

Note that the departure process D in (2.11) is somewhat more complicated than the workload process W because, unlike the workload process, the departure process cannot be represented directly as a function of the potential workload process X or the net-input process $C - S$. In general, the departure process cannot be represented directly in terms of X or $C - S$ because these processes cannot see the values of jumps in C and S that occur at the same time. Simultaneous jumps in C and S correspond to instants at which fluid arrives and some of it is instantaneously processed. The fluid that is instantaneously processed immediately upon arrival never affects the workload process. To obtain stochastic-process limits for the departure process, we will impose a condition to rule out such cancelling jumps in the limit processes associated with C and S . In particular, the departure process is considerably less complicated in the case of constant processing, as in (2.1).

We may also be interested in the *processing time* $T(t)$, i.e., the time required to process the work in the system at any time t , not counting any future input. For the processing time to correctly represent the actual processing time for the last particle of fluid in the queue, the fluid must be processed in the order of arrival. The processing time $T(t)$ is the first passage time to the level $W(t)$ by the future-available-processing process $\{S(t+u) - S(t) : u \geq 0\}$, i.e.,

$$T(t) \equiv \inf\{u \geq 0 : S(t+u) - S(t) \geq W(t)\}, \quad t \geq 0. \quad (2.12)$$

We can obtain an equivalent representation, involving a first passage time of the process S alone on the left in the infimum, if we use formula (2.9) for $W(t)$:

$$\begin{aligned} T(t) + t &= t + \inf\{u \geq 0 : S(t+u) - S(t) \geq X(t) + L(t) - U(t)\}, \\ &= \inf\{u \geq 0 : S(u) \geq W(0) + C(t) + L(t) - U(t)\}, \quad t \geq 0. \end{aligned} \quad (2.13)$$

In general, the processing time is relatively complicated, but in the common case of constant processing in (2.1), $T(t)$ is a simple modification of $W(t)$, namely,

$$T(t) = W(t)/\mu, \quad t \geq 0. \quad (2.14)$$

More generally, heavy-traffic limits also lead to such simplifications; see Section 5.9.2.

5.3. Unstable Queues

There are two main reasons queues experience congestion (which here means buildup of workload): First, the queue may be *unstable* (or overloaded); i.e., the input rate may exceed the output rate for an extended period of time, when there is ample storage capacity. Second, the queue may be stable, i.e., the long-run input rate may be less than the long-run output rate, but nevertheless short-run fluctuations produce temporary periods during which the input exceeds the output.

The unstable case tends to produce more severe congestion, but the stable case is more common, because systems are usually designed to be stable. Unstable queues typically arise in the presence of system failures. Since there is interest in system performance in the presence of failures, there is interest in the performance of unstable queues. For our discussion of unstable queues, we assume that there is unlimited storage capacity. We are interested in the buildup of congestion, which is described by the transient (or time-dependent) behavior of the queueing processes.

5.3.1. Fluid Limits for Fluid Queues

For unstable queues, useful insight can be gained from *fluid limits* associated with functional laws of large numbers (FLLN's). These stochastic-process limits are called fluid limits because the limit processes are deterministic functions of the form ct for some constant c . (More generally, with time-varying input and output rates, the limits could be deterministic functions of the form $\int_0^t r(s)ds$, $t \geq 0$, for some deterministic integrable function r .)

To express the FLLN's, we scale space and time both by n . As before, we use bold capitals to represent the scaled stochastic processes and associated limiting stochastic processes in the function space D . We use a hat to denote scaled stochastic processes with the fluid scaling (scaling space as well as time by n). Given the stochastic processes defined for the fluid-queue model in the previous section, form the associated scaled stochastic processes

$$\begin{aligned}\hat{\mathbf{C}}_n(t) &\equiv n^{-1}C(nt), \\ \hat{\mathbf{S}}_n(t) &\equiv n^{-1}S(nt), \\ \hat{\mathbf{X}}_n(t) &\equiv n^{-1}X(nt), \\ \hat{\mathbf{W}}_n(t) &\equiv n^{-1}W(nt), \\ \hat{\mathbf{L}}_n(t) &\equiv n^{-1}L(nt), \\ \hat{\mathbf{D}}_n(t) &\equiv n^{-1}D(nt),\end{aligned}$$

$$\hat{\mathbf{T}}_n(t) \equiv n^{-1}T(nt), \quad t \geq 0. \quad (3.1)$$

The continuous-mapping approach shows that FLLN's for C and S imply a joint FLLN for all the processes. As before, let $\mathbf{e}$ be the identity map, i.e., $\mathbf{e}(t) = t$, $t \geq 0$. Let $\mu \wedge \lambda \equiv \min\{\mu, \lambda\}$ and $\lambda^+ \equiv \max\{\lambda, 0\}$ for constants λ and μ .

We understand D to be the space $D([0, \infty), \mathbb{R})$, endowed with either the J_1 or the M_1 topology, as defined in Section 3.3. Since the limits are continuous deterministic functions, the J_1 and M_1 topologies here are equivalent to uniform convergence on compact subintervals. As in Section 3.3, we use D^k to denote the k -dimensional product space with the product topology; then $x_n \rightarrow x$, where $x_n \equiv (x_n^1, \dots, x_n^k)$ and $x \equiv (x^1, \dots, x^k)$, if and only if $x_n^i \rightarrow x^i$ for each i .

We first establish a functional weak law of large numbers (FWLLN), involving convergence in probability or, equivalently (because of the deterministic limit), convergence in distribution (see p. 27 of Billingsley (1999)). As indicated above, we restrict attention to the infinite-capacity model. It is easy to extend the results to the finite-capacity model, provided that the capacity is allowed to increase with n , as in Section 2.3.

Theorem 5.3.1. (FWLLN for the fluid queue) *In the infinite-capacity fluid-queue model, if $\hat{\mathbf{C}}_n \Rightarrow \lambda \mathbf{e}$ and $\hat{\mathbf{S}}_n \Rightarrow \mu \mathbf{e}$ in (D, M_1) , where $0 < \mu < \infty$ and $\hat{\mathbf{C}}_n$ and $\hat{\mathbf{S}}_n$ are given in (3.1), then*

$$(\hat{\mathbf{C}}_n, \hat{\mathbf{S}}_n, \hat{\mathbf{X}}_n, \hat{\mathbf{W}}_n, \hat{\mathbf{L}}_n, \hat{\mathbf{D}}_n, \hat{\mathbf{T}}_n) \Rightarrow (\lambda \mathbf{e}, \mu \mathbf{e}, (\lambda - \mu) \mathbf{e}, (\lambda - \mu)^+ \mathbf{e}, (\mu - \lambda)^+ \mathbf{e}, (\lambda \wedge \mu) \mathbf{e}, (\rho - 1)^+ \mathbf{e}) \quad (3.2)$$

in $(D, M_1)^7$ for $\rho \equiv \lambda/\mu$.

Proof. The single limits can be combined into joint limits because the limits are deterministic, by virtue of Theorem 11.4.5. So start with the joint convergence

$$(\hat{\mathbf{C}}_n, \hat{\mathbf{S}}_n, n^{-1}W(0)) \Rightarrow (\lambda \mathbf{e}, \mu \mathbf{e}, 0) \quad \text{in} \quad (D, M_1)^2 \times \mathbb{R}.$$

Since

$$\hat{\mathbf{X}}_n = \hat{\mathbf{C}}_n - \hat{\mathbf{S}}_n + n^{-1}W(0)$$

by (2.4), we can apply the continuous-mapping approach with addition, using the fact that addition on D^2 is measurable and continuous almost surely with respect to the limit process, to get the limit

$$\hat{\mathbf{X}}_n \Rightarrow \hat{\mathbf{X}} \equiv (\lambda - \mu) \mathbf{e}.$$

Specifically, we invoke Theorems 3.4.3 and 12.7.3 and Remark 12.7.1.

Then, because of (2.5) – (2.8), we can apply the simple continuous-mapping theorem, Theorem 3.4.1, with the reflection map to get

$$\hat{\mathbf{W}}_n \Rightarrow \hat{\mathbf{W}} \equiv \phi(\hat{\mathbf{X}}) = (\lambda - \mu)^+ \mathbf{e}$$

and

$$\hat{\mathbf{L}}_n \Rightarrow \hat{\mathbf{L}} \equiv \psi_L(\hat{\mathbf{X}}) = (\mu - \lambda)^+ \mathbf{e} ,$$

drawing on Theorems 13.5.1, 13.4.1 and 14.8.5. Then, by (2.11), we can apply the continuous-mapping approach with addition again to obtain $\hat{\mathbf{D}}_n \Rightarrow \hat{\mathbf{D}} = (\lambda \wedge \mu) \mathbf{e}$. Finally, by (2.13),

$$n^{-1}T(nt) + t = \inf\{u \geq 0 : n^{-1}S(nu) \geq n^{-1}(C(nt) + L(nt) + W(0))\} \quad (3.3)$$

or, in more compact notation,

$$\hat{\mathbf{T}}_n + \mathbf{e} = \hat{\mathbf{S}}_n^{-1} \circ (\hat{\mathbf{C}}_n + \hat{\mathbf{L}}_n + n^{-1}W(0)) . \quad (3.4)$$

Hence, we can again apply the continuous-mapping approach, this time with the inverse and composition functions. As with addition used above, these functions as maps from D and $D \times D$ to D are measurable and continuous almost surely with respect to the deterministic, continuous, strictly increasing limits. Specifically, by Corollary 13.6.4 and Theorem 13.2.1, we obtain

$$\hat{\mathbf{T}}_n + \mathbf{e} \Rightarrow \mu^{-1} \mathbf{e} \circ (\lambda \mathbf{e} + (\mu - \lambda)^+ \mathbf{e}) = (\rho \vee 1) \mathbf{e} ,$$

so that

$$\hat{\mathbf{T}}_n \Rightarrow (\rho - 1)^+ \mathbf{e} ,$$

as claimed. By Theorem 11.4.5, all limits can be joint. ■

From Theorem 5.3.1, we can characterize stable queues and unstable queues by the conditions $\lambda \leq \mu$ and $\lambda > \mu$, respectively, where λ and μ are the translation constants in the limits for the input process C and the available-processing process S . Equivalently, we can use the *traffic intensity* ρ , defined as

$$\rho \equiv \lambda/\mu . \quad (3.5)$$

From the relatively crude fluid-limit perspective, there is no congestion if $\rho \leq 1$; i.e., Theorem 5.3.1 implies that $\hat{\mathbf{W}}_n \Rightarrow 0\mathbf{e}$ if $\rho \leq 1$. On the other hand, if $\rho > 1$, then the workload tends to grow linearly at rate $\lambda - \mu$. Consistent with intuition, the fluid limits suggest using a simple deterministic analysis to describe congestion in unstable queues. When a queue is unstable

for a significant time, the relatively simple deterministic analysis may capture the dominant congestion effect. The same reasoning applies to queues with time-dependent input and output rates that are unstable for substantial periods of time. See Oliver and Samuel (1962), Newell (1982) and Hall (1991) for discussions of direct deterministic analysis of the congestion in queues.

Ordinary weak laws of large numbers (WLLN's), such as

$$t^{-1}W(t) \Rightarrow (\lambda - \mu)^+ \quad \text{in } \mathbb{R} \quad \text{as } t \rightarrow \infty ,$$

follow immediately from the FWLLN's in Theorem 5.3.1 by applying the continuous-mapping approach with the projection map, which maps a function x into $x(1)$. We could not obtain these WLLN's or the stronger FWLLN's in Theorem 5.3.1 if we assumed only ordinary WLLN's for C and S , i.e., if we had started with limits such as

$$t^{-1}C(t) \Rightarrow \lambda \quad \text{in } \mathbb{R} \quad \text{as } t \rightarrow \infty ,$$

because we needed to exploit the continuous-mapping approach in the function space D . We cannot go directly from a WLLN to a FWLLN, because a FWLLN is strictly stronger than a WLLN.

However, we can obtain functional strong laws of large numbers (FSLLN's) starting from ordinary strong laws of large numbers (SLLN's), because a SLLN implies a corresponding FSLLN; see Theorem 3.2.1 and Corollary 3.2.1 in the Internet Supplement. To emphasize that point, we now state the SLLN version of Theorem 5.3.1. Once we go from the SLLN's for C and S to the FSLLN's, the proof is the same as for Theorem 5.3.1.

Theorem 5.3.2. (FSLLN for the fluid queue) *In the infinite-capacity fluid-queue model, if*

$$t^{-1}C(t) \rightarrow \lambda \quad \text{and} \quad t^{-1}S(t) \rightarrow \mu \quad \text{in } \mathbb{R} \quad \text{w.p.1} \quad \text{as } t \rightarrow \infty ,$$

for $0 < \mu < \infty$, then

$$\begin{aligned} (\hat{\mathbf{C}}_n, \hat{\mathbf{S}}_n, \hat{\mathbf{X}}_n, \hat{\mathbf{W}}_n, \hat{\mathbf{L}}_n, \hat{\mathbf{D}}_n, \hat{\mathbf{T}}_n) \rightarrow \\ (\lambda \mathbf{e}, \mu \mathbf{e}, (\lambda - \mu) \mathbf{e}, (\lambda - \mu)^+ \mathbf{e}, (\mu - \lambda)^+ \mathbf{e}, (\lambda \wedge \mu) \mathbf{e}, (\rho - 1)^+ \mathbf{e}) \end{aligned} \quad (3.6)$$

w.p.1 in $(D, M_1)^7$ for ρ in (3.5).

5.3.2. Stochastic Refinements

We can also employ stochastic-process limits to obtain a more detailed description of congestion in unstable queues. These stochastic-process limits yield *stochastic refinements to the fluid limits* in Theorems 5.3.1 and 5.3.2 above. For the stochastic refinements, we introduce new scaled stochastic processes:

$$\begin{aligned}
\mathbf{C}_n(t) &\equiv c_n^{-1}(C(nt) - \lambda nt), \\
\mathbf{S}_n(t) &\equiv c_n^{-1}(S(nt) - \mu nt), \\
\mathbf{X}_n(t) &\equiv c_n^{-1}(X(nt) - (\lambda - \mu)nt), \\
\mathbf{W}_n(t) &\equiv c_n^{-1}(W(nt) - (\lambda - \mu)^+ nt), \\
\mathbf{L}_n(t) &\equiv c_n^{-1}(L(nt) - (\mu - \lambda)^+ nt), \\
\mathbf{D}_n(t) &\equiv c_n^{-1}(D(nt) - (\lambda \wedge \mu)nt), \\
\mathbf{T}_n(t) &\equiv c_n^{-1}(T(nt) - (\rho - 1)^+ nt), \quad t \geq 0.
\end{aligned} \tag{3.7}$$

As in the last chapter, the space scaling constants will be assumed to satisfy $c_n \rightarrow \infty$ and $n/c_n \rightarrow \infty$ as $n \rightarrow \infty$. The space-scaling constants will usually be a power, i.e., $c_n = n^H$ for $0 < H < 1$, but we allow other possibilities. In the following theorem we only discuss the cases $\rho < 1$ and $\rho > 1$. The more complex boundary case $\rho = 1$ is covered as a special case of results in the next section. Recall that D^k is the product space with the product topology; here we let the component space $D \equiv D^1$ have either the J_1 or the M_1 topology.

Since the limit processes $\mathbf{C}$ and $\mathbf{S}$ below may now have discontinuous sample paths, we need an extra condition to apply the continuous-mapping approach with addition. The extra condition depends on random sets of discontinuity points; e.g.,

$$Disc(\mathbf{S}) \equiv \{t : \mathbf{S}(t) \neq \mathbf{S}(t-)\},$$

where $x(t-)$ is the left limit of the function x in D (see Section 12.2). The random set of common discontinuity points of $\mathbf{C}$ and $\mathbf{S}$ is $Disc(\mathbf{C}) \cap Disc(\mathbf{S})$. The jump in $\mathbf{S}$ associated with a discontinuity at t is $\mathbf{S}(t) - \mathbf{S}(t-)$. The required extra condition is somewhat weaker for the M_1 topology than for the J_1 topology.

Theorem 5.3.3. (FCLT's for the stable and unstable fluid queues) *In the infinite-capacity fluid queue, suppose that $c_n \rightarrow \infty$ and $c_n/n \rightarrow 0$ as $n \rightarrow \infty$.*

Suppose that

$$(\mathbf{C}_n, \mathbf{S}_n) \Rightarrow (\mathbf{C}, \mathbf{S}) \quad \text{in } D^2, \quad (3.8)$$

where D^2 has the product topology with the topology on D^1 being either J_1 or M_1 , $\mathbf{C}_n$ and $\mathbf{S}_n$ are defined in (3.7) and

$$P(\mathbf{C}(0) = \mathbf{S}(0) = 0) = 1. \quad (3.9)$$

If the topology is J_1 , assume that $\mathbf{C}$ and $\mathbf{S}$ almost surely have no common discontinuities. If the topology is M_1 , assume that $\mathbf{C}$ and $\mathbf{S}$ almost surely have no common discontinuities with jumps of common sign.

(a) If $\rho < 1$ and $\mathbf{C} - \mathbf{S}$ has no positive jumps, then

$$\begin{aligned} (\mathbf{C}_n, \mathbf{S}_n, \mathbf{X}_n, \mathbf{W}_n, \mathbf{L}_n, \mathbf{D}_n) &\Rightarrow \\ (\mathbf{C}, \mathbf{S}, \mathbf{C} - \mathbf{S}, 0\mathbf{e}, \mathbf{S} - \mathbf{C}, \mathbf{C}) &\end{aligned} \quad (3.10)$$

in D^6 with the same topology.

(b) If $\rho > 1$, then

$$\begin{aligned} (\mathbf{C}_n, \mathbf{S}_n, \mathbf{X}_n, \mathbf{W}_n, \mathbf{L}_n, \mathbf{D}_n) &\Rightarrow \\ (\mathbf{C}, \mathbf{S}, \mathbf{C} - \mathbf{S}, \mathbf{C} - \mathbf{S}, 0\mathbf{e}, \mathbf{S}) &\end{aligned} \quad (3.11)$$

in D^6 with the same topology.

Proof. Paralleling the proof of Theorem 5.3.1 above, we start by applying condition (3.8) and Theorem 11.4.5 to obtain the joint convergence

$$(\mathbf{C}_n, \mathbf{S}_n, c_n^{-1}W(0)) \Rightarrow (\mathbf{C}, \mathbf{S}, 0) \quad \text{in } D^2 \times \mathbb{R}.$$

Then, as before, we apply the continuous mapping approach with addition, now invoking the conditions on the discontinuities of $\mathbf{C}$ and $\mathbf{S}$, to get

$$(\mathbf{C}_n, \mathbf{S}_n, \mathbf{X}_n, c_n^{-1}W(0)) \Rightarrow (\mathbf{C}, \mathbf{S}, \mathbf{C} - \mathbf{S}, 0) \quad \text{in } D^3 \times \mathbb{R}. \quad (3.12)$$

For the M_1 topology, we apply Theorems 3.4.3 and 12.7.3 and Remark 12.7.1. For J_1 , we apply the J_1 analog of Corollary 12.7.1; see Remark 12.6.2.

The critical step is treating $\mathbf{W}_n$. For that purpose, we apply Theorem 13.5.2, for which we need to impose the extra condition that $C - S$ have no positive jumps in part (a). We also use condition (3.9), but it can be weakened. We can use the Skorohod representation theorem, Theorem 3.2.2, to carry out the argument for individual sample paths.

The limit for $\mathbf{L}_n$ in part (a) then follows from (2.6), again exploiting the continuous-mapping approach with addition. The limits for $\mathbf{L}_n$ in part

(b) follows from Theorem 13.4.4, using (2.8) and condition (3.9). We can apply the convergence-together theorem, Theorem 11.4.7, to get limits for the scaled departure process $\mathbf{D}_n$. If $\lambda < \mu$, then

$$d_t(\mathbf{D}_n, \mathbf{C}_n) \leq \|\mathbf{D}_n - \mathbf{C}_n\|_t \leq \|c_n^{-1}W(0) - \mathbf{W}_n\|_t \Rightarrow 0$$

by (2.11), where d_t and $\|\cdot\|$ are the J_1 (or M_1) and uniform metrics for the time interval $[0, t]$, as in equations (3.2) and (3.1) of Section 3.3. If $\lambda > \mu$, then

$$d_t(\mathbf{D}_n, \mathbf{S}_n) \leq \|\mathbf{D}_n - \mathbf{S}_n\|_t \leq \|\mathbf{L}_n\|_t \Rightarrow 0$$

by (2.11). ■

The obvious sufficient condition for the limit processes $\mathbf{C}$ and $\mathbf{S}$ to almost surely have no discontinuities with jumps of common sign is to have no common discontinuities at all. For that, it suffices for $\mathbf{C}$ and $\mathbf{S}$ to be independent processes without any fixed discontinuities; i.e., $\mathbf{C}$ has no fixed discontinuities if $P(t \in \text{Disc}(\mathbf{C})) = 0$ for all t .

With the J_1 topology, the conclusion can be strengthened to the strong SJ_1 topology instead of the product J_1 topology, but that is not true for M_1 ; see Remark 9.3.1 and Example 14.5.1.

When $\rho < 1$, we not only obtain the zero fluid limit $\hat{\mathbf{W}}_n \Rightarrow 0\mathbf{e}$ in Theorem 5.3.1, but we also obtain the zero limit $\mathbf{W}_n \Rightarrow 0\mathbf{e}$ in Theorem 5.3.3 (a) with the refined scaling in (3.7), provided that $C - S$ has no positive jumps. However, if $C - S$ has positive jumps, then the scaled workload process $\mathbf{W}_n$ fails to be uniformly negligible. That shows the impact of jumps in the limit process.

Under extra conditions, we get a limit for $\mathbf{T}_n$ jointly with the limit in Theorem 5.3.3.

Theorem 5.3.4. (FCLT for the processing time) *Let the conditions of Theorem 5.3.3 hold. If the topology is J_1 , assume that S has no positive jumps.*

(a) *If $\rho < 1$, then jointly with the limit in (3.10)*

$$\mathbf{T}_n \Rightarrow 0\mathbf{e}$$

in D with the same topology.

(b) *Suppose that $\rho > 1$. If the topology is J_1 , assume that $\mathbf{C}$ and $\mathbf{S} \circ \rho\mathbf{e}$ almost surely have no common discontinuities. If the topology is M_1 , assume that $\mathbf{C}$ and $\mathbf{S} \circ \rho\mathbf{e}$ almost surely have no common discontinuities with jumps of common sign. Then jointly with the limit in (3.11)*

$$\mathbf{T}_n \Rightarrow \mu^{-1}(\mathbf{C} - \mathbf{S} \circ \rho\mathbf{e})$$

in D with the same topology.

Proof. We can apply Theorem 13.7.4 to treat $\mathbf{T}_n$, starting from (3.3) and (3.4). If $\lambda > \mu$, then

$$(n/c_n)(\hat{\mathbf{S}}_n - \mu\mathbf{e}, \hat{\mathbf{C}}_n + \hat{\mathbf{L}}_n + n^{-1}W(0) - \lambda\mathbf{e}) \Rightarrow (\mathbf{S}, \mathbf{C}) , \quad (3.13)$$

because $\mathbf{L}_n \Rightarrow 0\mathbf{e}$ and $n^{-1}W(0) \Rightarrow 0$. If $\lambda < \mu$, then

$$(n/c_n)(\hat{\mathbf{S}}_n - \mu\mathbf{e}, \hat{\mathbf{C}}_n + \hat{\mathbf{L}}_n + n^{-1}W(0) - \mu\mathbf{e}) \Rightarrow (\mathbf{S}, \mathbf{S}) , \quad (3.14)$$

because, by (2.6),

$$d_t(\mathbf{C}_n + \mathbf{L}_n + c_n^{-1}W(0), \mathbf{S}_n) \leq \|\mathbf{L}_n + \mathbf{X}_n\|_t = \|\mathbf{W}_n\|_t \Rightarrow 0 .$$

We can apply Theorem 13.7.4 to obtain limits for $\mathbf{T}_n$ jointly with the other limits because

$$\begin{aligned} \mathbf{T}_n &= (n/c_n)(\hat{\mathbf{T}}_n - (\rho - 1)^+\mathbf{e}) \\ &= (n/c_n)(\hat{\mathbf{S}}_n^{-1} \circ \hat{\mathbf{Z}}_n - (\rho \vee 1)\mathbf{e}) \\ &= (n/c_n)(\hat{\mathbf{S}}_n^{-1} \circ \hat{\mathbf{Z}}_n - \mu^{-1}\mathbf{e} \circ (\lambda \vee \mu)\mathbf{e}) \end{aligned}$$

for appropriate $\mathbf{Z}_n$ (specified in (3.13) and (3.14) above), where $n/c_n \rightarrow \infty$ as $n \rightarrow \infty$. Theorem 13.7.4 requires condition (3.9) for $\mathbf{S}$. ■

We regard the unstable case $\rho > 1$ as the case of primary interest for a single model. When $\rho > 1$, Theorem 5.3.3 (b) concludes that $W(t)$ obeys the same FCLT as $X(t)$. In a long time scale, the amount of reflection is negligible. Thus we obtain the approximation

$$W(t) \approx (\lambda - \mu)t + c_n\mathbf{X}(t/n) \quad (3.15)$$

for the workload, where $\mathbf{X} = \mathbf{C} - \mathbf{S}$. In the common setting of Donsker's theorem, $c_n = n^{1/2}$ and $\mathbf{X} = \sigma_X\mathbf{B}$, where $\mathbf{B}$ is standard Brownian motion. In that special case, (3.15) becomes

$$\begin{aligned} W(t) &\approx (\lambda - \mu)t + n^{1/2}\sigma_X\mathbf{B}(t/n) \\ &\approx N((\lambda - \mu)t, \sigma_X^2 t) . \end{aligned} \quad (3.16)$$

In this common special case, the stochastic refinement of the LLN shows that the workload obeys a CLT and, thus, the workload $W(t)$ should be approximately normally distributed with mean equal to the fluid limit $(\lambda - \mu)t$ and standard deviation proportional to $\sqrt{t}$, with the variability parameter given explicitly. With heavy tails or strong dependence (or both), but still with finite mean, the stochastic fluctuations about the mean will be greater, as is made precise by the stochastic-process limits.

Remark 5.3.1. *Implications for queues in series.* Part (a) of Theorem 5.3.3 has important implications for queues in series: If the first of two queues is stable with $\rho < 1$, then the departure process D at the first queue obeys the same FCLT as the input process C at that first queue. Thus, if we consider a heavy-traffic limit for the second queue (either because the second queue is unstable or because we consider a sequence of models for the second queue with the associated sequence of traffic intensities at the second queue approaching the critical level for stability, as in the next section), then the heavy-traffic limit at the second queue depends on the first queue only through the input stochastic process at that first queue. In other words, the heavy-traffic behavior of the second queue is the same as if the first queue were not even there. We obtain more general and more complicated heavy-traffic stochastic-process limits for the second queue only if we consider a sequence of models for both queues, and simultaneously let the sequences of traffic intensities at both queues approach the critical levels for stability, which puts us in the setting of Chapter 14. For further discussion, see Example 9.9.1, Chapter 14 and Karpelovich and Kreinin (1994). ■

In this section we have seen how heavy-traffic stochastic-process limits can describe the congestion in an unstable queue. We have considered the relatively elementary case of constant input and output rates. Variations of the same approach apply to queues with time-varying input and output rates; see Massey and Whitt (1994a), Mandelbaum and Massey (1995), Mandelbaum, Massey and Reiman (1998) and Chapter 9 of the Internet Supplement.

5.4. Heavy-Traffic Limits for Stable Queues

We now want to establish nondegenerate heavy-traffic stochastic-process limits for stochastic processes in stable fluid queues (where the long-run input rate is less than the maximum potential output rate). (With a finite storage capacity, the workload will of course remain bounded even if the long-run input rate exceeds the output rate.)

The first heavy-traffic limits for queues were established by Kingman (1961, 1962, 1965). The treatment here is in the spirit of Iglehart and Whitt (1970a, b) and Whitt (1971a), although those papers focused on standard queueing models, as considered here in Chapters 9 and 10. An early heavy-traffic limit for finite-capacity queues was established by Kennedy

(1973). See Whitt (1974b) and Borovkov (1976, 1984) for background on early heavy-traffic limits.

In order to establish the heavy-traffic stochastic-process limits for stable queues, we consider a sequence of models indexed by a subscript n , where the associated sequence of traffic intensities $\{\rho_n : n \geq 1\}$ converges to 1, the critical level for stability, as $n \rightarrow \infty$. We have in mind the case in which the traffic intensities approach 1 from below, denoted by $\rho_n \uparrow 1$, but that is not strictly required. For each n , there is a cumulative-input process C_n , an available-processing process S_n , a storage capacity K_n with $0 < K_n \leq \infty$ and an initial workload $W_n(0)$ satisfying $0 \leq W_n(0) \leq K_n$. As before, we make no specific structural or stochastic assumptions about the stochastic processes C_n and S_n , so we have very general models. A more detailed model for the input is considered in Chapter 8.

To have the traffic intensity well defined in our setting, we assume that the limits

$$\lambda_n \equiv \lim_{t \rightarrow \infty} t^{-1} C_n(t) \quad (4.1)$$

and

$$\mu_n \equiv \lim_{t \rightarrow \infty} t^{-1} S_n(t) \quad (4.2)$$

exist w.p.1 for each n . We call λ_n the *input rate* and μ_n the *maximum potential output rate* for model n . (The actual output rate is the input rate minus the overflow rate.) Then the traffic intensity in model n is

$$\rho_n \equiv \lambda_n / \mu_n . \quad (4.3)$$

We will be letting $\rho_n \rightarrow 1$ as $n \rightarrow \infty$.

Given the basic model elements above, we can construct the potential-workload processes $\{X_n(t) : t \geq 0\}$, the workload processes $\{W_n(t) : t \geq 0\}$, the upper-barrier regulator (overflow) processes $\{U_n(t) : t \geq 0\}$, the lower-barrier regulator processes $\{L_n(t) : t \geq 0\}$ and the departure processes $\{D_n(t) : t \geq 0\}$ as described in Sections 5.2.

We now form associated scaled processes. We could obtain fluid limits in this setting, paralleling Theorems 5.3.1 and 5.3.2, but they add little beyond the previous results. Hence we go directly to the generalizations of Theorem 5.3.3. We scale the processes as in (3.7), but now we have processes and translation constants for each n . Let

$$\begin{aligned} \mathbf{C}_n(t) &\equiv c_n^{-1}(C_n(nt) - \lambda_n nt) , \\ \mathbf{S}_n(t) &\equiv c_n^{-1}(S_n(nt) - \mu_n nt) , \\ \mathbf{X}_n(t) &\equiv c_n^{-1}X_n(nt) , \end{aligned}$$

$$\begin{aligned}
\mathbf{W}_n(t) &\equiv c_n^{-1} W_n(nt) , \\
\mathbf{U}_n(t) &\equiv c_n^{-1} U_n(nt) , \\
\mathbf{L}_n(t) &\equiv c_n^{-1} L_n(nt) , \quad t \geq 0 .
\end{aligned} \tag{4.4}$$

For the scaling constants, we have in mind $\lambda_n \rightarrow \lambda$ and $\mu_n \rightarrow \mu$ as $n \rightarrow \infty$, where $0 < \lambda < \infty$ and $0 < \mu < \infty$, with $c_n \rightarrow \infty$ and $n/c_n \rightarrow \infty$ as $n \rightarrow \infty$. As in Section 2.3, the upper barrier must grow as $n \rightarrow \infty$; specifically, we require that $K_n = c_n K$.

Our key assumption is a joint limit for $\mathbf{C}_n$ and $\mathbf{S}_n$ in (4.4). When there are limits for $\mathbf{C}_n$ and $\mathbf{S}_n$ with the translation terms involving λ_n and μ_n , the w.p.1 limits in (4.1) and (4.2) usually hold too, but (4.1) and (4.2) are actually not required. However, convergence in probability in (4.1) and (4.2) follows directly as a consequence of the convergence in distribution assumed below. Hence it is natural for the limits in (4.1) and (4.2) to hold as well.

Let (ϕ_K, ψ_U, ψ_L) be the reflection map mapping a potential-workload process X into the triple (W, U, L) , as defined in Section 5.2. Here is the general heavy-traffic stochastic-process limit for stable fluid queues. It follows directly from the continuous-mapping approach using addition and reflection.

Theorem 5.4.1. (general heavy-traffic limit for stable fluid queues) *Consider a sequence of fluid queues indexed by n with capacities K_n , $0 < K_n \leq \infty$, general cumulative-input processes $\{C_n(t) : t \geq 0\}$ and general cumulative-available-processing processes $\{S_n(t) : t \geq 0\}$. Suppose that $K_n = c_n K$, $0 < K \leq \infty$, $0 \leq W_n(0) \leq K_n$,*

$$(c_n^{-1} W_n(0), \mathbf{C}_n, \mathbf{S}_n) \Rightarrow (W'(0), \mathbf{C}, \mathbf{S}) \quad \text{in } \mathbb{R} \times D^2 \tag{4.5}$$

for $\mathbf{C}_n$ and $\mathbf{S}_n$ in (4.4), where the topology on D^2 is the product topology with the topology on D^1 being either J_1 or M_1 , $c_n \rightarrow \infty$, $c_n/n \rightarrow 0$ and $\lambda_n - \mu_n \rightarrow 0$, so that

$$\eta_n \equiv n(\lambda_n - \mu_n)/c_n \rightarrow \eta , \tag{4.6}$$

where $-\infty < \eta < \infty$. If the topology is J_1 , suppose that almost surely $\mathbf{C}$ and $\mathbf{S}$ have no common discontinuities. If the topology is M_1 , suppose that almost surely $\mathbf{C}$ and $\mathbf{S}$ have no common discontinuities with jumps of common sign. Then, jointly with the limit in (4.5),

$$(\mathbf{X}_n, \mathbf{W}_n, \mathbf{U}_n, \mathbf{L}_n) \Rightarrow (\mathbf{X}, \mathbf{W}, \mathbf{U}, \mathbf{L}) \tag{4.7}$$

in D^4 with the same topology, where

$$\mathbf{X}(t) = W'(0) + \mathbf{C}(t) - \mathbf{S}(t) + \eta t, \quad t \geq 0. \quad (4.8)$$

and

$$(\mathbf{W}, \mathbf{U}, \mathbf{L}) \equiv (\phi_K(\mathbf{X}), \psi_U(\mathbf{X}), \psi_L(\mathbf{X})) \quad (4.9)$$

with (ϕ_K, ψ_U, ψ_L) being the reflection map associated with capacity K .

Proof. Note that

$$\mathbf{X}_n = c_n^{-1} W_n(0) + \mathbf{C}_n - \mathbf{S}_n + \eta_n \mathbf{e}, \quad (4.10)$$

where $\mathbf{e}(t) \equiv t$ for $t \geq 0$. Thus, just as in Theorems 5.3.1 and 5.3.3 above, we can apply the continuous-mapping approach starting from the joint convergence

$$(c_n^{-1} W_n(0), \mathbf{C}_n, \mathbf{S}_n, \eta_n \mathbf{e}) \Rightarrow (W'(0), \mathbf{C}, \mathbf{S}, \eta \mathbf{e}) \quad (4.11)$$

in $\mathbb{R} \times D^3$, which follows from (4.5), (4.6) and Theorem 11.4.5. We apply the continuous mapping theorem, Theorem 3.4.3, with addition to get $\mathbf{X}_n \Rightarrow \mathbf{X}$. (Alternatively, we could use the Skorohod representation theorem, Theorem 3.2.2.) We use the fact that addition is measurable and continuous almost surely with respect to the limit process, by virtue of the assumption about the discontinuities of $\mathbf{C}$ and $\mathbf{S}$. Specifically, for M_1 we apply Remark 12.7.1 and Theorem 12.7.3. For J_1 we apply the analog of Corollary 12.7.1; see Remark 12.6.2. Finally, we obtain the desired limit in (4.7) because

$$(\mathbf{W}_n, \mathbf{U}_n, \mathbf{L}_n) = (\phi_K(\mathbf{X}_n), \psi_U(\mathbf{X}_n), \psi_L(\mathbf{X}_n))$$

for all n . We apply the simple continuous-mapping theorem, Theorem 3.4.1, with the reflection maps, using the continuity established in Theorems 13.5.1 and 14.8.5. ■

Just as in Theorem 5.3.3, with the J_1 topology the conclusion holds in the strong SJ_1 topology as well as the product J_1 topology. As before, the conditions on the common discontinuities of $\mathbf{C}$ and $\mathbf{S}$ hold if $\mathbf{C}$ and $\mathbf{S}$ are independent processes without fixed discontinuities.

In the standard heavy-traffic applications, in addition to (4.6), we have $\lambda_n < \mu_n$, $\mu_n \rightarrow \mu$ for $0 < \mu < \infty$, $\lambda_n - \mu_n \rightarrow 0$ and $\rho_n \equiv \lambda_n/\mu_n \uparrow 1$. However, we can have non-heavy-traffic limits by having $\lambda_n n/c_n \rightarrow a > 0$ and $\mu_n n/c_n \rightarrow b > 0$, so that $c = a - b$ and $\rho_n \equiv \lambda_n/\mu_n \rightarrow a/b$, where a/b can be any positive value. Nevertheless, the heavy-traffic limit with $\rho_n \uparrow 1$ is the principal case.

We discuss heavy-traffic stochastic-process limits for the departure process and the processing time in Section 5.9. Before discussing the implications of Theorem 5.4.1, we digress to put the heavy-traffic limits in perspective with other asymptotic methods.

Remark 5.4.1. *The long tradition of asymptotics.* Given interest in the distribution of the workload $W(t)$, we perform the heavy-traffic limit, allowing $\rho_n \uparrow 1$ as $n \rightarrow \infty$ in a sequence of models index by n , to obtain simplified expressions for the cdf $P(W(t) > x)$ and the distribution of the entire process $\{W(t) : t \geq 0\}$. We describe the resulting approximation in the Brownian case in Section 5.7 below. To put the heavy-traffic limit in perspective, we should view it in the broader context of asymptotic methods: For general mathematical models, there is a long tradition of applying asymptotic methods to obtain tractable approximations; e.g., see Bender and Orszag (1978), Bleistein and Handelsman (1986) and Olver (1974). In this tradition are the heavy-traffic approximations and asymptotic expansions obtained by Knessl and Tier (1995, 1998) using singular perturbation methods.

For stochastic processes, it is customary to perform asymptotics. We usually simplify by letting $t \rightarrow \infty$: Under regularity conditions, we obtain $W(t) \Rightarrow W(\infty)$ as $t \rightarrow \infty$ and then we focus on the limiting steady-state cdf $P(W(\infty) > x)$. (Or, similarly, we look for a stationary distribution of the process $\{W(t) : t \geq 0\}$.) This asymptotic step is so common that it is often done without thinking. See Asmussen (1987), Baccelli and Bremaud (1994) and Borovkov (1976) for supporting theory for basic queueing processes. See Bramson (1994a,b), Baccelli and Foss (1994), Dai (1994), Meyn and Down (1994) and Borovkov (1998) for related stability results for queueing networks and more general processes.

Given a steady-state cdf $P(W(\infty) > x)$, we may go further and let $x \rightarrow \infty$ to find the steady-state tail-probability asymptotics. As noted in Section 2.4.1, a common case for a queue with unlimited waiting space is the exponential tail:

$$P(W(\infty) > x) \sim \alpha e^{-\eta x} \quad \text{as } x \rightarrow \infty ,$$

which yields the simple exponential approximation

$$P(W(\infty) > x) \approx \alpha e^{-\eta x}$$

for all x not too small; e.g., see Abate, Choudhury and Whitt (1994b, 1995).

With exponential tail-probability asymptotics, the key quantity is the asymptotic decay rate η . Since α is much less important than η , we may

ignore α (i.e., let $\alpha = 1$), which corresponds to exploiting weaker large-deviation asymptotics of the form

$$\log P(W(\infty) > x) \sim -\eta x \quad \text{as } x \rightarrow \infty ;$$

e.g., see Glynn and Whitt (1994) and Schwartz and Weiss (1995).

The large deviations limit is associated with the concept of effective bandwidths used for admission control in communication networks; see Berger and Whitt (1998a,b), Chang and Thomas (1995), Choudhury, Lucantoni and Whitt (1996), de Veciana, Kesidis and Walrand (1995), Kelly (1996) and Whitt (1993b). The idea is to assign a deterministic quantity, called the effective bandwidth, to represent how much capacity a source will require. New sources are then admitted if the sum of the effective bandwidths does not exceed the available bandwidth.

We will also consider tail-probability asymptotics applied to the steady-state distribution of the heavy-traffic limit process. We could instead consider heavy-traffic limits after establishing tail-probability asymptotics. It is significant that the two iterated limits often agree: Often the heavy-traffic asymptotics for η as $\rho \uparrow 1$ matches the asymptotics as first $t \rightarrow \infty$ and then $x \rightarrow \infty$ in the heavy-traffic limit process; see Abate and Whitt (1994b) and Choudhury and Whitt (1994). More generally, Majewski (2000) has shown that large-deviation and heavy traffic limits for queues can be interchanged. The large-deviation and heavy-traffic views are directly linked by moderate-deviations limits, which involve a different scaling, including heavy traffic ($\rho_n \uparrow 1$); see Puhalskii (1999) and Wischik (2001b).

However, as noted in Section 2.4.1, other asymptotic forms are possible for queueing processes. We often have

$$P(W(\infty) > x) \sim \alpha x^{-\beta} e^{-\eta x} \quad \text{as } x \rightarrow \infty , \quad (4.12)$$

for non-zero β ; e.g., see Abate and Whitt (1997b), Choudhury and Whitt (1996) and Duffield (1997). Moreover, even other asymptotic forms are possible; e.g., see Flatto (1997).

With heavy-tailed distributions, we usually have a power tail, i.e., (4.12) holds with $\eta = 0$:

$$P(W(\infty) > x) \sim \alpha x^{-\beta} \quad \text{as } x \rightarrow \infty .$$

When the steady-state distribution of the workload in a queue has a power tail, the heavy-traffic theory usually is consistent; i.e., the heavy-traffic limits usually capture the relevant tail asymptotics; see Section 8.5. For more on

power-tail asymptotics, see Abate, Choudhury and Whitt (1994a), Duffield and O'Connell (1995), Boxma and Dumas (1998), Sigman (1999), Jelenković (1999, 2000), Likhanov and Mazumdar (2000), Whitt (2000c) and Zwart (2000, 2001).

With the asymptotic form in (4.12), numerical transform inversion can be used to calculate the asymptotic constants η , β and α from the Laplace transform, as shown in Abate, Choudhury, Lucantoni and Whitt (1995) and Choudhury and Whitt (1996). When $\eta = 0$, we can transform the distribution into one with $\eta > 0$ to perform the computation; see Section 5 of Abate, Choudhury and Whitt (1994a) and Section 3 of Abate and Whitt (1997b). See Abate and Whitt (1996, 1999a,b,c) for ways to construct heavy-tailed distributions with tractable Laplace transforms.

And there are many other kinds of asymptotics that can be considered. For example, with queueing networks, we can let the size of the network grow; e.g., see Whitt (1984e, 1985c), Kelly (1991), Vvedenskaya et al. (1996), Mitzenmacher (1996), and Turner (1998) ■

5.5. Heavy-Traffic Scaling

A primary reason for establishing the heavy-traffic stochastic-process limit for stable queues in the previous section is to generate approximations for the workload stochastic process in a stable fluid-queue model. However, it is not exactly clear how to do this, because in applications we have one given queueing system, not a sequence of queueing systems. The general idea is to regard our given queueing system as the n^{th} queueing system in the sequence of queueing systems, but what should the value of n be?

The standard way to proceed is to choose n so that the traffic intensity ρ_n in the sequence of systems matches the actual traffic intensity in the given system. That procedure makes sense because the traffic intensity ρ is a robust first-order characterization of the system, not depending upon the stochastic fluctuations about long-term rates. As can be seen from (4.1) – (4.3) and Theorems 5.3.1 and 5.3.2, the traffic intensity appears in the fluid scaling. Thus, it is natural to think of the heavy-traffic stochastic-process limit as a way to capture the second-order variability effect beyond the traffic intensity ρ .

In controlled queueing systems, it may be necessary to solve an optimization problem to determine the relevant traffic intensity. Then the traffic intensity can not be regarded as given, but instead must be derived; see Harrison (2000, 2001a,b). After deriving the traffic intensity, we may

proceed with further heavy-traffic analysis. Here we assume that the traffic intensity has been determined.

If we decide to choose n so that the traffic intensity ρ_n matches the given traffic intensity, then it is natural to index the models by the traffic intensity ρ from the outset, and then consider the limit as $\rho \uparrow 1$ (with $\uparrow$ indicating convergence upward from below). In this section we show how we can index the queueing models by the traffic intensity ρ instead of an arbitrary index n . We also discuss the applied significance of the scaling of space and time in heavy-traffic stochastic-process limits. We focus on the general fluid model considered in the last two sections, but the discussion applies to even more general models.

5.5.1. The Impact of Scaling Upon Performance

Let $W_\rho(t)$ denote the workload at time t in the infinite-capacity fluid-queue model with traffic intensity ρ . Let $c(\rho)$ and $b(\rho)$ denote the functions that scale space and time, to be identified in the next subsection. Then the scaled workload process is

$$\mathbf{W}_\rho(t) \equiv c(\rho)^{-1} W_\rho(b(\rho)t) \quad t \geq 0. \quad (5.1)$$

The heavy-traffic stochastic-process limit can then be expressed as

$$\mathbf{W}_\rho \Rightarrow \mathbf{W} \quad \text{in } (D, M_1) \quad \text{as } \rho \uparrow 1, \quad (5.2)$$

where $D \equiv D([0, \infty), \mathbb{R})$ and $\{\mathbf{W}(t) : t \geq 0\}$ is the limiting stochastic process. In the limits we consider, $c(\rho) \uparrow \infty$ and $b(\rho) \uparrow \infty$ as $\rho \uparrow 1$. Thus, the heavy-traffic stochastic-process limit provides a macroscopic view of uncertainty.

Given the heavy-traffic stochastic-process limit for the workload process in (5.2), the natural approximation is obtained by replacing the limit by approximate equality in distribution; i.e.,

$$c(\rho)^{-1} W_\rho(b(\rho)t) \approx \mathbf{W}(t), \quad t \geq 0,$$

or, equivalently, upon moving the scaling terms to the right side,

$$W_\rho(t) \approx c(\rho) \mathbf{W}(b(\rho)^{-1}t), \quad t \geq 0, \quad (5.3)$$

where $\approx$ means approximately equal to in distribution (as stochastic processes).

We first discuss the applied significance of the two scaling functions $c(\rho)$ and $b(\rho)$ appearing in (5.1) and (5.3). Then, afterwards, we show how to identify these scaling functions for the fluid-queue model.

The scaling functions $c(\rho)$ and $b(\rho)$ provide important insight into queueing performance. The space-scaling factor $c(\rho)$ is relatively easy to interpret: The workload process (for times not too small) tends to be of order $c(\rho)$ as $\rho \uparrow 1$. The time-scaling factor $b(\rho)$ is somewhat more subtle: The workload process tends to make significant changes over time scales of order $b(\rho)$ as $\rho \uparrow 1$. Specifically, the change in the workload process, when adjusted for space scaling, from time $t_1 b(\rho)$ to time $t_2 b(\rho)$ is approximately characterized (for suitably high ρ) by the change in the limit process $\mathbf{W}$ from time t_1 to time t_2 .

Consequently, over time intervals of length less than $b(\rho)$ the workload process tends to remain unchanged. Specifically, if we consider the change in the workload process W_ρ from time $t_1 b(\rho)$ to time $t_2(\rho)$, where $t_2(\rho) > t_1 b(\rho)$ but $t_2(\rho)/b(\rho) \rightarrow 0$ as $\rho \uparrow 1$, and if the limit process $\mathbf{W}$ is almost surely continuous at time t_1 , then we conclude from the heavy-traffic limit in (5.2) that the relative change in the workload process over the time interval $[t_1 b(\rho), t_2(\rho)]$ is asymptotically negligible as ρ increases.

On the other hand, over time intervals of length greater than $b(\rho)$, the workload process W_ρ tends to approach its equilibrium steady-state distribution (assuming that both $\mathbf{W}(t)$ and $W_\rho(t)$ approach steady-state limits as $t \rightarrow \infty$). Specifically, when $t_2(\rho) > t_1 b(\rho)$ and $t_2(\rho)/b(\rho) \rightarrow \infty$ as $\rho \uparrow 1$, the workload process at time $t_2(\rho)$ tends to be in steady state, independent of its value at time $t_1 b(\rho)$. Thus, if we are considering the workload process over the time interval $[t_1 b(\rho), t_2(\rho)]$, we could use steady-state distributions to describe the distribution of $W_\rho(t_2(\rho))$, ignoring initial conditions at time $t_1 b(\rho)$. (In that step, we assume that $\mathbf{W}(t)$ approaches a steady-state distribution as $t \rightarrow \infty$, independent of initial conditions.) Thus, under regularity conditions, the time scaling in the heavy-traffic limit reveals the rate of convergence to steady state, as a function of the traffic intensity.

The use of steady-state distributions tends to be appropriate only over time intervals of length greater than $b(\rho)$. Since $b(\rho) \uparrow \infty$ as $\rho \uparrow 1$, transient (time-dependent) analysis becomes more important as ρ increases. Fortunately, the heavy-traffic stochastic-process limits provide a basis for analyzing the approximate transient behavior of the workload process as well as the approximate steady-state behavior. As indicated above, the change in the workload process (when adjusted for space scaling) between times $t_1 b(\rho)$ and $t_2 b(\rho)$ is approximately characterized by the change in the limit process $\mathbf{W}$ from time t_1 to time t_2 . Fortunately, the limit processes often

are sufficiently tractable that we can calculate such transient probabilities.

Remark 5.5.1. *Relaxation times.* The approximate time for a stochastic process to approach its steady-state distribution is called the *relaxation time*; e.g., see Section III.7.3 of Cohen (1982). The relaxation time can be defined in a variety of ways, but it invariably is based on the limiting behavior as $t \rightarrow \infty$ for fixed ρ . In the relatively nice light-tailed and weak-dependent case, it often can be shown, under regularity conditions, that

$$E[f(W_\rho(t))] - E[f(W_\rho(\infty))] \sim g(t, \rho)e^{-t/r(\rho)} \quad \text{as } t \rightarrow \infty, \quad (5.4)$$

for various real-valued functions f , with the functions g and r in general depending upon f . The standard asymptotic form for the second-order term g is $g(t, \rho) \sim c(\rho)$ or $g(t, \rho) \sim c(\rho)t^{\beta(\rho)}$ as $t \rightarrow \infty$. When (5.4) holds with such a g , $r(\rho)$ is called the relaxation time. Of course, a stochastic process that starts away from steady state usually does not reach steady state in finite time. Instead, it gradually approaches steady state in a manner such as described in (5.4). More properly, we should interpret $1/r(\rho)$ as the rate of approach to steady state.

With light tails and weak dependence, we usually have

$$r(\rho)/b(\rho) \rightarrow c \quad \text{as } \rho \uparrow 1,$$

where c is a positive constant; i.e., the heavy-traffic time-scaling usually reveals the asymptotic form (as $\rho \uparrow 1$) of the relaxation time.

However, with heavy tails and strong dependence, the approach to steady state is usually much slower than in (5.4); see Asmussen and Teugels (1996) and Mikosch and Nagaev (2000). In these other settings, as well as in the light-tailed weak-dependent case, the time scaling in the heavy-traffic limit usually reveals the asymptotic form (as $\rho \uparrow 1$) of the approach to steady state. Thus, the heavy-traffic time scaling can provide important insight into the rate of approach to steady state. With heavy tails and strong dependence, the heavy-traffic limits show that transient analysis becomes more important. ■

5.5.2. Identifying Appropriate Scaling Functions

We now consider how to identify appropriate scaling functions $b(\rho)$ and $c(\rho)$ in (5.1). We can apply the general stochastic-process limit in Theorem 5.4.1 to determine appropriate scaling functions. Specifically, the scaling functions $b(\rho)$ and $c(\rho)$ depend on the input rates λ_n , the output rates μ_n

and the space-scaling factors c_n appearing in Theorem 5.4.1. The key limit is (4.6), which determines the drift η of the unreflected limit process $\mathbf{X}$.

To cover most cases of practical interest, we make *three additional assumptions* about the scaling as a function of n in (4.4): First, we assume that the space scaling is by a simple power. Specifically, we assume that

$$c_n \equiv n^H \quad \text{for } 0 < H < 1. \quad (5.5)$$

(See Section 4.2 for discussion about the possible scaling functions.) We need the condition on the exponent H in (5.5) in order to have $c_n \rightarrow \infty$ and $c_n/n \rightarrow 0$ as $n \rightarrow \infty$, as assumed in Theorem 5.4.1.

Second, we assume that the translation terms λ_n and μ_n in (4.4) converge to finite positive limits as $n \rightarrow \infty$. In view of condition (4.6) in Theorem 5.4.1, it suffices to assume only that

$$\mu_n \rightarrow \mu \quad \text{as } n \rightarrow \infty, \quad (5.6)$$

where $0 < \mu < \infty$.

Third, we assume that the basic limit in (4.6) holds with $\eta < 0$. That implies that the traffic intensities ρ_n are less than 1 for all n sufficiently large. Now, if we combine (4.6), (5.5) and (5.6) (and divide by μ_n in (4.6)), we obtain the condition

$$n^{1-H}(1 - \rho_n) \rightarrow \zeta \equiv -\eta/\mu > 0 \quad (5.7)$$

for $0 < \zeta < \infty$. From (5.7), we obtain the associated limit

$$n(1 - \rho_n)^{1/(1-H)} \rightarrow \zeta^{1/(1-H)} \quad \text{as } n \rightarrow \infty \quad (5.8)$$

or, equivalently,

$$n \sim \left(\frac{\zeta}{1 - \rho_n} \right)^{\frac{1}{1-H}} \quad \text{as } n \rightarrow \infty. \quad (5.9)$$

Thus the *canonical forms of the scaling functions* are

$$b(\rho) \equiv n \equiv \left(\frac{\zeta}{1 - \rho} \right)^{\frac{1}{1-H}} \quad (5.10)$$

and

$$c(\rho) \equiv n^H \equiv \left(\frac{\zeta}{1 - \rho} \right)^{\frac{H}{1-H}} \quad (5.11)$$

for $\zeta = -\eta/\mu$ as in (5.7).

To summarize, when the net-input process and potential-workload process satisfies a FCLT with time scaling by n and space scaling by n^H , the associated scaled workload processes, as functions of the traffic intensity ρ , have a heavy-traffic limit with the time-scaling function in (5.10) and space-scaling function in (5.11); i.e., as functions of ρ , the *time-scaling exponent* is $1/(1 - H)$ and the *space-scaling exponent* is $H/(1 - H)$.

The initial space-scaling exponent H (the Hurst parameter) depends on the burstiness; see Chapter 4. As the burstiness increases, H increases. Of course, the standard case, considered in most heavy-traffic limits for queues, is $H = 1/2$. The standard case with $H = 1/2$ occurs with Donsker's theorem and its variants with weak dependence and light tails, as discussed in Sections 4.3 and 4.4. Since $H = 1/2$ is the standard case, it is also the reference case. Values of H with $1/2 < H < 1$ indicate greater burstiness associated with heavy tails or strong positive dependence (or both). Values of H with $0 < H < 1/2$ are associated with strong negative dependence, as might occur with strong traffic shaping, e.g., scheduling.

From (5.10) and (5.11), we see that the scaling functions $b(\rho)$ and $c(\rho)$ increase rapidly as $H \uparrow 1$ for ρ near 1. Indeed, the scaling exponents increase as H increases from 0 toward 1. To make that important point clear, we display the two scaling exponents for a range of H values in Table 5.1.

H	time-scaling exponent $1/(1 - H)$	space-scaling exponent $H/(1 - H)$
1/101	101/100	1/100
1/11	11/10	1/10
1/5	5/4	1/4
1/3	3/2	1/2
1/2	2	1
2/3	3	2
4/5	5	4
10/11	11	10
100/101	101	100

Table 5.1: The time-scaling and space-scaling exponents as a function of the Hurst parameter H .

Since H increases as the burstiness increases, we see that increased burstiness leads to greater scaling functions $c(\rho)$ and $b(\rho)$ for any given traffic intensity ρ . The larger value of $c(\rho)$ shows that the buffer content is

likely to be larger (or that one needs larger buffers to avoid overflow). The larger values of $b(\rho)$ show that the time scales for statistical regularity are longer. When there is larger burstiness, transient analysis becomes more important in contrast to steady-state analysis.

From a practical engineering perspective, the analysis of the heavy-traffic scaling functions $b(\rho)$ and $c(\rho)$ indicates that, when exceptional variability is a possibility in a queueing setting, attention should be focused on the space-scaling exponent H for the net-input process as well as the traffic intensity ρ . Second-order refinements are provided by the constant ζ appearing in (5.7), (5.10) and (5.11) and the limit process $\mathbf{W}$ appearing in (5.2) and (5.3).

5.6. Limits as the System Size Increases

In this section we see how heavy-traffic stochastic-process limits for stable fluid queues change as the system size increases. The heavy-traffic limits thus show how performance scales as the system size increases. We will see that *the performance impact depends on the way that the system size increases*. We start with a base infinite-capacity fluid queue for which there is a heavy-traffic stochastic-process limit. We assume that there is a limit for the potential-workload processes of the form $\mathbf{X}_n \Rightarrow \mathbf{X}$, where

$$\mathbf{X}_n(t) \equiv n^{-H} X_n(nt), \quad t \geq 0, \quad (6.1)$$

for $0 < H < 1$ and

$$\mathbf{X}(t) \equiv \eta t + \mathbf{Y}(t), \quad t \geq 0, \quad (6.2)$$

with $\{\mathbf{Y}(t) : t \geq 0\}$ being H -self-similar, i.e.,

$$\{\mathbf{Y}(ct) : t \geq 0\} \stackrel{d}{=} \{c^H \mathbf{Y}(t) : t \geq 0\} \quad (6.3)$$

as in (2.5) in Section 4.2. Of course, there is a corresponding heavy-traffic stochastic-process limit for the workload process,

$$\mathbf{W}_n \Rightarrow \mathbf{W} \equiv \phi(\mathbf{X}),$$

where

$$\mathbf{W}_n \equiv \phi(\mathbf{X}_n).$$

It will be convenient to focus on the potential-workload processes $\mathbf{X}_n$ instead of the workload processes $\mathbf{W}_n$. We will focus on the scale factor σ when the limit process has the representation $\mathbf{X} \equiv \eta \mathbf{e} + \sigma \mathbf{Y}$. For fixed η and $\mathbf{Y}$, the associated reflection $\{\mathbf{W}(t) : t \geq 0\}$ tends to be increasing in

σ (in a stochastic sense). For example, if $\mathbf{Y}$ is standard Brownian motion and $\eta < 0$, then the steady-state quantity $\mathbf{W}(\infty)$ has mean $\sigma^2/2|\eta|$; see (7.13) below. More generally, σ serves as a quantitative measure of the variability (for fixed $\mathbf{Y}$). The general principle is: *Increased variability in the potential workload process leads to larger workloads*, where “larger” is measured appropriately, e.g., by the mean or by a form of stochastic order.

We consider three ways to make the system larger: scaling space, scaling time and creating independent replicas. Let the *size-increase factor* be a positive integer m . We *scale space* (make it larger) by considering $m\mathbf{X}_n$; we *scale time* (make it faster) by considering $\mathbf{X}_n \circ m\mathbf{e}$; and we *create independent replicas* by considering $\mathbf{X}_{n,1} + \cdots + \mathbf{X}_{n,m}$, where $\mathbf{X}_{n,1}, \dots, \mathbf{X}_{n,m}$ are m IID copies of the original stochastic processes $\mathbf{X}_n$.

For communication network applications, it is useful to think of constant deterministic processing, whose rate is being increased by a factor m . Scaling space then amounts to making the files or packets m times bigger to match the increased capacity. Scaling time amounts to sending the same input m times faster. Creating independent replicas means superposing (adding) m independent sources, each distributed as the original one. (We will be considering heavy-traffic limits for superposition input processes further in later chapters; see Sections 8.7.1, 9.4 and 9.8.)

In manufacturing, scaling space can also occur. Scaling space occurs in batching and unbatching; e.g., see Sections 8.5 and 9.3 of Hopp and Spearman (1996).

When we scale space, the limit process is

$$m\mathbf{X} = m\eta\mathbf{e} + m\mathbf{Y} . \quad (6.4)$$

When we scale time, the limit process is

$$\begin{aligned} \mathbf{X} \circ m\mathbf{e} &= m\eta\mathbf{e} + \mathbf{Y} \circ m\mathbf{e} \\ &\stackrel{d}{=} m\eta\mathbf{e} + m^H\mathbf{Y} . \end{aligned} \quad (6.5)$$

When we create independent replicas, the limit process is

$$\sum_{i=1}^m \mathbf{X}_i = m\eta\mathbf{e} + \sum_{i=1}^m \mathbf{Y}_i . \quad (6.6)$$

The rate of the limit process increases by the same factor m in all three cases, but the impact on the stochastic component, characterized by the stochastic process $\mathbf{Y}$, is different for the three methods. Scaling time by m produces smaller stochastic fluctuations than scaling space by m , in the

sense that the scale factors before $\mathbf{Y}$ in (6.4) and (6.5) are ordered: $m^H < m$. The advantage of time scaling over space scaling increases as H decreases (when the variability is smaller).

The impact of creating independent replicas depends on the properties of the stochastic process $\mathbf{Y}$. If $\mathbf{Y}$ is a Lévy process (has stationary and independent increments), then a concatenation of independent versions is equivalent to a longer version, i.e.,

$$\sum_{i=1}^m \mathbf{Y}_i \stackrel{d}{=} \mathbf{Y} \circ m\mathbf{e} . \quad (6.7)$$

Thus, if $\mathbf{Y}$ is a Lévy process, creating independent replicas is equivalent to scaling time, which we have seen produces better performance than scaling space.

On the other hand, suppose that $\mathbf{Y}$ is fractional Brownian motion (FBM), the principal example of a non-Lévy limit process in Chapter 4. Since FBM is not a Lévy process, (6.7) does not hold. When $\mathbf{Y}$ is FBM, both $\mathbf{Y}$ and $\sum_{i=1}^m \mathbf{Y}_i$ are zero-mean Gaussian processes. For zero-mean Gaussian processes, it is natural to focus on the variances. With independent replicas, the variance is

$$\text{Var} \sum_{i=1}^m \mathbf{Y}_i(t) = m(\text{Var} \mathbf{Y}(t)), \quad t \geq 0 . \quad (6.8)$$

In contrast, with time scaling, because of the H -self-similarity, the variance is

$$\text{Var} \mathbf{Y}(mt) = \text{Var}(m^H \mathbf{Y}(t)) = m^{2H}(\text{Var} \mathbf{Y}(t)) . \quad (6.9)$$

Hence, the variance with independent replicas is less than, equal to or greater than the variance with time scaling, respectively, when $H > 1/2$, $H = 1/2$ or $H < 1/2$.

More generally, we can compare all three methods using the variance when $\mathbf{Y}(t)$ has finite variance. Using the H -self-similarity of $\mathbf{Y}$, we obtain

$$\begin{aligned} \text{Var}(m\mathbf{Y}(t)) &= m^2(\text{Var} \mathbf{Y}(t)), \\ \text{Var} \mathbf{Y}(mt) &= m^{2H}(\text{Var} \mathbf{Y}(t)), \\ \text{Var} \sum_{i=1}^m \mathbf{Y}_i(t) &= m(\text{Var} \mathbf{Y}(t)) . \end{aligned} \quad (6.10)$$

For $H < 1/2$, time scaling produces least variability; for $H > 1/2$, independent replicas produces least variability.

It is interesting to compare one large system (increased by factor m) to m separate independent systems, distributed as the original one. We say that there is *economy of scale* when the workload in the single large system tends to be smaller than the sum of the workloads in the separate systems. With finite variances, there is economy of scale when the ratio of the standard deviation to the mean is decreasing in m . From (6.10), we see that there is economy of scale with time scaling and independent replicas, but not with space scaling. For communication networks, the economy of scale associated with independent replicas is often called the *multiplexing gain*, i.e., the gain in efficiency from statistical multiplexing (combining independent sources). See Smith and Whitt (1981) for stochastic comparisons demonstrating the economy of scale in queueing systems. See Chapters 8 and 9 for more discussion.

Example 5.6.1. Brownian motion. Suppose that $\mathbf{X} = \eta \mathbf{e} + \sigma \mathbf{B}$, where $\eta < 0$, $\sigma > 0$ and $\mathbf{B}$ is standard Brownian motion. As noted above, the associated RBM has steady-state mean $\sigma^2/2|\eta|$. With space scaling, time scaling and creating independent replicas, the steady-state mean of the RBM's become

$$m\sigma^2/2|\eta|, \quad \sigma^2/2|\eta| \quad \text{and} \quad \sigma^2/2|\eta| ,$$

respectively. Thus, with space scaling, the steady-state mean is the same as the total steady-state mean in m separate systems. Otherwise, the steady-state mean is less by the factor m . ■

In this section we have considered three different ways that the fluid queue can get larger. We have shown that the three different ways have different performance implications. It is important to realize, however, that in applications the situation may be more complicated. For example, a computer can be made larger by adding processors, but there invariably are limitations that prevent the maximum potential output rate from being proportional to the number of processors as the number of processors increases.

If the jobs are processed one at a time, then we must exploit *parallel processing*, i.e., the processors must share the processing of each job. However, usually a proportion of each job cannot be parallelized. Thus, with parallel processing, the capacity tends to increase nonlinearly with the number of processors; the marginal gain in capacity tends to be decreasing in m ; e.g., see Amdahl (1967) and Chapters 5-7 and 14 of Gunther (1998). With deterministic processing, our analysis would still apply, provided that we interpret m as the actual increase in processing rate.

Even if we can accurately estimate the effective processing rate, there remain difficulties in applying the analysis in this section, because with parallel processing, it may not be appropriate to regard the processing as deterministic. It then becomes difficult to determine how the available-processing process S and its FCLT should change with m .

5.7. Brownian Approximations

In this section we apply the general heavy-traffic stochastic-process limits in Section 5.4 to establish Brownian heavy-traffic limits for fluid queues. In particular, under extra assumptions (corresponding to light tails and weak dependence), the limit for the normalized cumulative-input process will be a zero-drift Brownian motion (BM) and the limit for the normalized workload process will be a reflected Brownian motion (RBM), usually with negative drift.

The general heavy-traffic stochastic-process limits in Section 5.4 also generate non-Brownian approximations corresponding to the non-Brownian FCLT's in Chapter 4, but we do not discuss them here. We discuss approximations associated with stable Lévy motion and fractional Brownian approximations in Chapter 8.

Since Brownian motion has continuous sample paths and the reflection map maps continuous functions into continuous functions, RBM also has continuous sample paths. However, unlike Brownian motion, RBM does not have independent increments. But RBM is a Markov process. As a (well-behaved) Markov process with continuous sample paths, RBM is a diffusion process.

Harrison (1985) provides an excellent introduction to Brownian motion and “Brownian queues,” showing how they can be analyzed using martingales and the Ito stochastic calculus. Other good introductions to Brownian motion and diffusion processes are Glynn (1990), Karatzas and Shreve (1988) and Chapter 15 of Karlin and Taylor (1981). Borodin and Salminen (1996) provide many Brownian formulas. Additional properties of RBM are contained in Abate and Whitt (1987a-b, 1988a-d).

5.7.1. The Brownian Limit

If $\mathbf{B}$ is a standard Brownian motion, then $\{y + \eta t + \sigma \mathbf{B}(t) : t \geq 0\}$ is a Brownian motion with *drift* η , *diffusion coefficient* (or variance coefficient) σ^2 and initial position y . We have the following elementary application of Section 5.4.

Theorem 5.7.1. (general RBM limit) *Suppose that the conditions of Theorem 5.4.1 are satisfied with $W'(0) = y$, $c_n = \sqrt{n}$ and $(\mathbf{C}, \mathbf{S})$ two-dimensional zero-drift Brownian motion with covariance matrix*

$$\Sigma = \begin{pmatrix} \sigma_C^2 & \sigma_{C,S}^2 \\ \sigma_{C,S}^2 & \sigma_S^2 \end{pmatrix}. \quad (7.1)$$

Then the conclusions of Theorems 5.4.1, 5.9.1 and 5.9.3 (b) hold with

$$(\mathbf{W}, \mathbf{U}, \mathbf{L}) \equiv (\phi_K(\mathbf{X}), \psi_U(\mathbf{X}), \psi_L(\mathbf{X}))$$

being reflected Brownian motion, i.e.,

$$\mathbf{X}(t) \stackrel{d}{=} y + \eta t + \sigma_X \mathbf{B}(t) \quad (7.2)$$

for standard Brownian motion $\mathbf{B}$, drift coefficient η in (4.6) and diffusion coefficient

$$\sigma_X^2 = \sigma_C^2 + \sigma_S^2 - 2\sigma_{C,S}^2. \quad (7.3)$$

Proof. Under the assumption on $(\mathbf{C}, \mathbf{S})$, $\mathbf{C} - \mathbf{S}$ is a zero-drift Brownian motion with diffusion coefficient σ_X^2 in (7.3). ■

As indicated in Section 5.5, we can also index the queueing systems by the traffic intensity ρ and let $\rho \uparrow 1$. With $n = \zeta^2/(1-\rho)^2$ as in (5.10), the heavy-traffic limit becomes

$$\{\zeta^{-1}(1-\rho)W_\rho(t\zeta^2/(1-\rho)^2) : t \geq 0\} \Rightarrow \phi_K(\tilde{\mathbf{X}}) \quad \text{as } \rho \uparrow 1, \quad (7.4)$$

where W_ρ is the workload process in model ρ , which has output rate μ and traffic intensity ρ , and

$$\tilde{\mathbf{X}}(t) \stackrel{d}{=} y - \zeta\mu t + \mathbf{B}(\sigma_X^2 t), \quad t \geq 0, \quad (7.5)$$

with $\mathbf{B}$ being a standard Brownian motion. The capacity in model ρ is $K_\rho = \zeta K/(1-\rho)$.

We have freedom in the choice of the parameter ζ . If we let

$$\zeta = \sigma_X^2/\mu, \quad (7.6)$$

and rescale time by replacing t by t/σ_X^2 , then the limit in (7.4) can be expressed as

$$\{\sigma_X^{-2}\mu(1-\rho)W_\rho(t\sigma_X^2/\mu^2(1-\rho)^2) : t \geq 0\} \Rightarrow \phi_K(\mathbf{X}) \quad (7.7)$$

where $\mathbf{X}$ is canonical Brownian motion with drift coefficient -1 and variance coefficient 1 , plus initial position y , i.e.,

$$\{\mathbf{X}(t) : t \geq 0\} \stackrel{d}{=} \{y - t + \mathbf{B}(t) : t \geq 0\} .$$

That leads to the *Brownian approximation*

$$\{W_\rho(t) : t \geq 0\} \approx \{\sigma_X^2 \mu^{-1} (1 - \rho)^{-1} \phi_K(\mathbf{X}) (\mu^2 (1 - \rho)^2 t / \sigma_X^2) : t \geq 0\} , \quad (7.8)$$

where $\mathbf{X}$ is again canonical Brownian motion.

Remark 5.7.1. *The impact of variability* The Brownian limit and the Brownian approximation provide insight into the way variability in the basic stochastic processes C and S affect queueing performance. In the heavy-traffic limit, the stochastic behavior of the processes C and S , beyond their rates λ and μ , affect the Brownian approximation solely via the single variance parameter σ_X^2 in (7.3), which can be identified from the CLT for $C - S$. For further discussion, see Section 9.6.1. ■

We now show how the Brownian approximation applies to the steady-state workload.

5.7.2. The Steady-State Distribution.

The heavy-traffic limit in Theorem 5.7.1 does not directly imply that the steady-state distributions converge. Nevertheless, from (7.8), we obtain an approximation for the steady-state workload, namely,

$$W_\rho(\infty) \approx \frac{\sigma_X^2}{\mu(1 - \rho)} \phi_K(\mathbf{X})(\infty) . \quad (7.9)$$

Conditions for the convergence of steady-state distributions in heavy traffic have been established by Szczotka (1986, 1990, 1999).

We now give the steady-state distribution of RBM with two-sided reflection; see p. 90 of Harrison (1985). We are usually interested in the case of negative drift, but we allow positive drift as well when $K < \infty$.

Theorem 5.7.2. (steady-state distribution of RBM) *Let $\{\mathbf{W}(t) : t \geq 0\}$ be one-dimensional RBM with drift coefficient η , diffusion coefficient σ^2 , initial value y and two-sided reflection at 0 and K . Then*

$$\mathbf{W}(t) \Rightarrow \mathbf{W}(\infty) \quad \text{in } \mathbb{R} \quad \text{as } t \rightarrow \infty ,$$

where $\mathbf{W}(\infty)$ has pdf

$$f(x) \equiv \begin{cases} 1/K & \text{if } \eta = 0 \\ \frac{\theta e^{\theta x}}{e^{\theta K} - 1} & \text{if } \eta \neq 0, \end{cases} \quad (7.10)$$

with mean

$$E\mathbf{W}(\infty) = \begin{cases} K/2, & \text{if } \eta = 0 \\ \frac{K}{1 - e^{-\theta K}} - \frac{1}{\theta} & \text{if } \eta \neq 0 \end{cases} \quad (7.11)$$

for

$$\theta \equiv 2\eta/\sigma^2 \quad (7.12)$$

Note that the steady-state distribution of RBM in (7.10) depends only on the two parameters θ in (7.12) and K . The steady-state distribution is uniform in the zero-drift case; the steady-state distribution is an exponential distribution with mean $-\theta^{-1} = \sigma^2/2|\eta|$, conditional on being in the interval $[0, K]$, when $\eta < 0$ and $\theta < 0$; $K - \mathbf{W}(\infty)$ has an exponential distribution with mean $\theta^{-1} = \sigma^2/2\eta$, conditional on being in the interval $[0, K]$, when $\eta > 0$ and $\theta > 0$. Without the upper barrier at K , a steady-state distribution exists if and only if $\eta < 0$, in which case it is the exponential distribution with mean $-\theta^{-1}$ obtained by letting $K \rightarrow \infty$ in (7.10). As K gets large, the tails of the exponential distributions rapidly become negligible so that

$$E\mathbf{W}(\infty) \approx \begin{cases} |\theta|^{-1} & \text{if } \eta < 0 \\ K - |\theta|^{-1} & \text{if } \eta > 0. \end{cases} \quad (7.13)$$

Let us now consider the approximation indicated by the limit. Since $n^{-1/2}\mathbf{W}_n(nt) \Rightarrow \mathbf{W}(t)$, we use the approximations

$$\mathbf{W}_n(t) \approx \sqrt{n}\mathbf{W}(t/n) \quad (7.14)$$

and

$$\mathbf{W}_n(\infty) \approx \sqrt{n}\mathbf{W}(\infty). \quad (7.15)$$

Thus, when $K = \infty$, the Brownian approximation for $W_\rho(\infty)$ is an exponential random variable with mean

$$E[W_\rho(\infty)] \approx \frac{\sigma_X^2}{2\mu(1 - \rho)}. \quad (7.16)$$

The RBM's $\phi_K(\tilde{\mathbf{X}})$ in (7.4) and $\phi_K(\mathbf{X})$ in (7.7) and (7.8) are the Brownian queues, which serve as the approximating models. From the approximations in (7.8) – (7.16), we see the impact upon queueing performance of the processes C and S in the heavy-traffic limit. In the heavy-traffic limit, the processes C and S affect performance through their rates $\lambda = \rho\mu$ and μ and through the variance parameter σ_X^2 , which depends on the elements of the covariance matrix Σ in (7.1) as indicated in (7.3).

Note in particular that the mean of RBM in (7.16) is directly proportional to the variability of $X = C - S$ through the variability parameter σ_X^2 in (7.3). The variability parameter σ_X^2 in turn is precisely the variance constant in the CLT for the net-input process $C - S$.

In (7.9)–(7.16) we have described the approximations for the steady-state workload distribution that follow directly from the heavy-traffic limit theorem in Theorem 5.7.1. It is also possible to modify or “refine” the approximations to satisfy other criteria. For example, extra terms that appear in known exact formulas for special cases, but which are negligible in the heavy-traffic limit, may be inserted. If the goal is to develop accurate numerical approximations, then it is natural to regard heavy-traffic limits as only one of the possible theoretical reference points. For the standard multi-server GI/G/s queue, for which the heavy-traffic limit is also RBM, heuristic refinements are discussed in Whitt (1982b, 1993a) and references therein.

For the fluid queue, an important reference case for which exact formulas are available is a single-source model with independent sequences of IID on times and off times (a special case of the model studied in Chapter 8). Kella and Whitt (1992b) show that the workload process and its steady-state distribution can be related to the virtual waiting time process in the standard GI/G/1 queue (studied here in Chapter 9). Relatively simple moment formulas are thus available in the M/G/1 special case. The steady-state workload distribution can be computed in the general GI/G/1 case using numerical transform inversion, following Abate, Choudhury and Whitt (1993, 1994a, 1999). Such computations were used to illustrate the performance of bounds for general fluid queues by Choudhury and Whitt (1997).

A specific way to generate refined approximations is to interpolate between light-traffic and heavy-traffic limits; see Burman and Smith (1983, 1986), Fendick and Whitt (1989), Reiman and Simon (1988, 1989), Reiman and Weiss (1989) and Whitt (1989b). Even though numerical accuracy can be improved by refinements, the direct heavy-traffic Brownian approximations remain appealing for their simplicity.

Example 5.7.1. *The M/G/1 steady state workload.* It is instructive to compare the approximations with exact values when we can determine them. For the standard M/G/1 queue with $K = \infty$, the mean steady-state workload has the simple exact formula

$$E[W_\rho(\infty)] = \frac{\rho\sigma_X^2}{2(1-\rho)} , \quad (7.17)$$

which differs from (7.16) only by the factor ρ in the numerator of (7.17) and the factor μ in the denominator of (7.16). First, in the M/G/1 model the workload process has constant output rate 1, so $\mu = 1$. Hence, the only real difference between (7.16) and (7.17) is the factor ρ in the numerator of (7.17), which approaches 1 in the heavy-traffic limit.

To elaborate, in the M/G/1 queue, the cumulative input $C(t)$ equals the sum of the service times of all arrivals in the interval $[0, t]$, i.e., the cumulative input is

$$C(t) \equiv \sum_{k=1}^{A(t)} V_k, \quad t \geq 0 ,$$

where $\{A(t) : t \geq 0\}$ is a rate- ν Poisson arrival process independent of the sequence $\{V_k : k \geq 1\}$ of IID service times, with V_1 having a general distribution with mean EV_1 . Thus, the traffic intensity is $\rho \equiv \nu EV_1$. The workload process is defined in terms of the net-input process $X(t) \equiv C(t) - t$ as described in Section 5.2.

The cumulative-input process is a special case of a renewal-reward process, considered in Section 7.4. Thus, by Theorem 7.4.1, if

$$\sigma_V^2 \equiv \text{Var} V_1 < \infty ,$$

then the cumulative-input process obeys a FCLT $\mathbf{C}_n \Rightarrow \mathbf{C}$ for $\mathbf{C}_n$ in (3.7) with translation constant $\lambda \equiv \rho$ and space-scaling function $c_n = n^{1/2}$. Then the limit process is $\sigma_C \mathbf{B}$, where $\mathbf{B}$ is standard Brownian motion and

$$\begin{aligned} \sigma_C^2 &= \nu\sigma_V^2 + \rho EV_1 \\ &= \rho EV_1 (c_V^2 + 1) , \end{aligned} \quad (7.18)$$

where c_V^2 is the squared coefficient of variation, defined by

$$c_V^2 \equiv \sigma_V^2 / (EV_1)^2 . \quad (7.19)$$

Therefore,

$$\sigma_X^2 = \sigma_C^2 = \rho EV_1 (c_V^2 + 1) . \quad (7.20)$$

With this notation, the exact formula for the mean steady-state workload in the M/G/1 queue is given in (7.17) above; e.g., see Chapter 5 of Kleinrock (1975). As indicated above, the approximation in (7.16) differs from the exact formula in (7.17) only by the factor ρ in the numerator of the exact formula, which of course disappears (becomes 1) in the heavy-traffic limit.

For the M/G/1 queue, it is known that

$$P(W_\rho(\infty) = 0) = 1 - \rho . \quad (7.21)$$

Thus, if we understand the approximation to be for the conditional mean $E[W_\rho(\infty) | W_\rho(\infty) > 0]$, then the approximation becomes exact. In general, however, the distribution of $W_\rho(\infty)$ is not exponential, so that the exponential distribution remains an approximation for the M/G/1 model, but the conditional distribution of $W(\infty)$ given that $W(\infty) > 0$ is exponential in the M/M/1 special case, in which the service-time distribution is exponential. ■

5.7.3. The Overflow Process

In practice it is also of interest to describe the overflow process. In a communication network, the overflow process describes lost packets. An important design criterion is to keep the packet loss rate below a specified threshold. The *loss rate* in model n is

$$\beta_n \equiv \lim_{t \rightarrow \infty} t^{-1} U_n(t) . \quad (7.22)$$

The limits in Theorems 5.4.1 and 5.7.1 show that, with the heavy-traffic scaling, the loss rate should be asymptotically negligible as $n \rightarrow \infty$. Specifically, since $n^{-1/2} U_n(nt) \Rightarrow \mathbf{U}(t)$ as $n \rightarrow \infty$, where $\mathbf{U}$ is the upper-barrier regulator process of RBM, the cumulative loss in the interval $[0, n]$ is of order $\sqrt{n}$, so that the loss rate should be of order $1/\sqrt{n}$ as $n \rightarrow \infty$. (Of course, this asymptotic form depends on having the upper barriers grow as $K_n = \sqrt{n}K$ and $\rho_n \rightarrow 1$.) More precisely, we approximate the loss rate β_n by

$$\beta_n \approx \beta / \sqrt{n} , \quad (7.23)$$

where

$$\beta \equiv \lim_{t \rightarrow \infty} t^{-1} \mathbf{U}(t) . \quad (7.24)$$

Note that approximation (7.23) involves an unjustified interchange of limits, involving $n \rightarrow \infty$ and $t \rightarrow \infty$.

Berger and Whitt (1992b) make numerical comparisons (based on exact numerical algorithms) showing how the Brownian approximation in (7.23) performs for finite-capacity queues. For very small loss rates, such as 10^{-9} , it is not possible to achieve high accuracy. (Systems with the same heavy-traffic limit may have loss rates varying from 10^{-4} to 10^{-15} .) Such very small probabilities tend to be captured better by large-deviations limits. For a simple numerical comparison, see Srikant and Whitt (2001). Overall, the Brownian approximation provides important insight. That is illustrated by the sensitivity analysis in Section 9 of Berger and Whitt (1992b).

More generally, the heavy-traffic stochastic-process limits support the approximation

$$\mathbf{U}_n(t) \approx \sqrt{n}\mathbf{U}(t/n), \quad t \geq 0, \quad (7.25)$$

where $\mathbf{U}$ is the upper-barrier regulator process of RBM. In order for the Brownian approximation for the overflow process in (7.25) to be useful, we need to obtain useful characterizations of the upper-barrier regulator process $\mathbf{U}$ associated with RBM. It suffices to describe one of the boundary regulation processes $\mathbf{U}$ and $\mathbf{L}$, because $\mathbf{L}$ has the same structure as $\mathbf{U}$ with a drift of the opposite sign. The rates of the process $\mathbf{L}$ and $\mathbf{U}$ are determined on p. 90 of Harrison (1985).

Theorem 5.7.3. (rates of boundary regulator processes) *The rates of the boundary regulator processes exist, satisfying*

$$\alpha \equiv \lim_{t \rightarrow \infty} \frac{\mathbf{L}(t)}{t} = \lim_{t \rightarrow \infty} \frac{E\mathbf{L}(t)}{t} = \begin{cases} \sigma^2/2K & \text{if } \eta = 0 \\ \frac{\eta}{e^{\theta K} - 1} & \text{if } \eta \neq 0 \end{cases} \quad (7.26)$$

and

$$\beta \equiv \lim_{t \rightarrow \infty} \frac{\mathbf{U}(t)}{t} = \lim_{t \rightarrow \infty} \frac{E\mathbf{U}(t)}{t} = \begin{cases} \sigma^2/2K & \text{if } \eta = 0 \\ \frac{\eta}{1 - e^{-\theta K}} & \text{if } \eta \neq 0. \end{cases} \quad (7.27)$$

It is important to note that the loss rate β depends upon the variance σ^2 , either directly (when $\eta = 0$) or via θ in (7.12). We can use regenerative analysis and martingales to further describe the Brownian boundary regulation processes $\mathbf{L}$ and $\mathbf{U}$; see Berger and Whitt (1992b) and Williams (1992). Let $T_{a,b}$ be the first passage time from level a to level b within $[0, K]$. Epochs at which RBM first hits 0 after first hitting K are regeneration points for the processes $\mathbf{L}$ and $\mathbf{U}$. Assuming that the RBM starts at 0, one regeneration

cycle is completed at time $T_{0,K} + T_{K,0}$. Of course, $\mathbf{L}$ increases only during $[0, T_{0,K}]$, while $\mathbf{U}$ increases only during $[T_{0,K}, T_{0,K} + T_{K,0}]$. We can apply regenerative analysis and the central limit theorem for renewal processes to show that the following limits exist

$$\alpha \equiv \lim_{t \rightarrow \infty} \frac{\mathbf{L}(t)}{t} = \lim_{t \rightarrow \infty} \frac{E\mathbf{L}(t)}{t} = \frac{E\mathbf{L}(T_{0,K} + T_{K,0})}{E(T_{0,K} + T_{K,0})} \quad (7.28)$$

$$\beta \equiv \lim_{t \rightarrow \infty} \frac{\mathbf{U}(t)}{t} = \lim_{t \rightarrow \infty} \frac{E\mathbf{U}(t)}{t} = \frac{E\mathbf{U}(T_{0,K} + T_{K,0})}{E(T_{0,K} + T_{K,0})} \quad (7.29)$$

$$\sigma_L^2 \equiv \lim_{t \rightarrow \infty} \frac{Var \mathbf{L}(t)}{t} \quad \text{and} \quad \sigma_U^2 \equiv \lim_{t \rightarrow \infty} \frac{Var \mathbf{U}(t)}{t}. \quad (7.30)$$

The parameters σ_L^2 and σ_U^2 in (7.30) are the *asymptotic variance parameters* of the processes $\mathbf{L}$ and $\mathbf{U}$. It is also natural to focus on the *normalized asymptotic variance parameters*

$$c_L^2 \equiv \sigma_L^2 / \alpha \quad \text{and} \quad c_U^2 \equiv \sigma_U^2 / \beta. \quad (7.31)$$

Theorem 5.7.4. (normalized asymptotic variance of boundary regulator processes) *The normalized asymptotic variance parameters in (7.31) satisfy*

$$\begin{aligned} c_U^2 &= c_L^2 = E \left[\left(\mathbf{L}(T_{0,K}) - \frac{(T_{0,K} + T_{K,0}) E\mathbf{L}(T_{0,K})}{E(T_{0,K} + T_{K,0})} \right)^2 / E\mathbf{L}(T_{0,K}) \right] \\ &= \begin{cases} 2K/3 & \text{if } \eta = 0 \\ \frac{2(1 - e^{2\theta K} + 4\theta K e^{\theta K})}{-\theta(1 - e^{\theta K})^2} & \text{if } \eta \neq 0 \end{cases} \end{aligned} \quad (7.32)$$

for $\theta \equiv 2\eta/\sigma^2$ as in (7.12).

In order to obtain the last line of (7.32) in Theorem 5.7.4, and for its own sake, we use an expression for the joint transform of $\mathbf{L}(T_{0,K})$ and $T_{0,K}$ from Williams (1992). Note that it suffices to let $\sigma^2 = 1$, because if $\sigma^2 > 0$ and $\mathbf{W}$ is a $(\eta/\sigma, 1)$ RBM on $[0, K/\sigma]$, then $\sigma\mathbf{W}$ is an (η, σ^2) -RBM on $[0, K]$.

Theorem 5.7.5. (joint distribution of key variables in the regenerative representation) *For $\sigma^2 = 1$ and all $s_1, s_2 \geq 0$,*

$$E[\exp(-s_1 \mathbf{L}(T_{0,K}) - s_2 T_{0,K})]$$

$$= \begin{cases} \frac{1}{1+s_1 K} & \text{if } \eta = 0, s_2 = 0 \\ \frac{1}{\cosh(\gamma K) + s_1 \gamma^{-1} \sinh(\gamma K)} & \text{if } \eta = 0, s_2 \neq 0 \\ \frac{e^{mK}}{\cos(\gamma K) + (s_1 + m) \gamma^{-1} \sinh(\gamma K)} & \text{if } \eta \neq 0, \end{cases} \quad (7.33)$$

where $\gamma = \sqrt{\eta^2 + 2s_2}$.

Since an explicit expression for the Laplace transform is available, we can exploit numerical transform inversion to calculate the joint probability distribution and the marginal probability distributions of $T_{0,K}$ and $\mathbf{L}(T_{0,K})$; see Abate and Whitt (1992a, 1995a), Choudhury, Lucantoni and Whitt (1994) and Abate, Choudhury and Whitt (1999).

Explicit expressions for the moments of $\mathbf{L}(T_{0,K})$ and $T_{0,K}$ can be obtained directly from Theorem 5.7.5.

Theorem 5.7.6. (associated moments of regenerative variables) *If $\eta = 0$ and $\sigma^2 = 1$, then*

$$\begin{aligned} ET_{0,K} &= K^2, \quad ET_{0,K}^2 = 5K^4/3, \\ E[\mathbf{L}(T_{0,K})] &= K, \quad E[\mathbf{L}(T_{0,K})^2] = 2K^2 \\ E[T_{0,K} \mathbf{L}(T_{0,K})] &= 5K^3/3. \end{aligned} \quad (7.34)$$

If $\eta \neq 0$ and $\sigma^2 = 1$, then

$$\begin{aligned} ET_{0,K} &= (e^{-2\eta K} - 1 + 2\eta K)/2\eta^2, \\ E[T_{0,K}^2] &= (e^{-4\eta K} + e^{-2\eta K} + 6\eta K e^{-2\eta K} + 2\eta^2 K^2 - 2)/2\eta^4, \\ E[\mathbf{L}(T_{0,K})] &= (1 - e^{-2\eta K})/2\eta, \\ E[\mathbf{L}(T_{0,K})^2] &= (1 - e^{-2\eta K})^2/2\eta^2, \\ E[T_{0,K} \mathbf{L}(T_{0,K})] &= (e^{-2\eta K} - 3\eta K e^{-2\eta K} - e^{-4\eta K} + \eta K)/2\eta^3. \end{aligned} \quad (7.35)$$

Fendick and Whitt (1998) show how a Brownian approximation can be used to help interpret loss measurements in a communication network.

5.7.4. One-Sided Reflection

Even nicer descriptions of RBM are possible when there is only one reflecting barrier at the origin (corresponding to an infinite buffer). Let $\mathbf{R} \equiv \{\mathbf{R}(t; \eta, \sigma^2, x) : t \geq 0\}$ denote RBM with one reflecting barrier at the

origin, i.e., $\mathbf{R} = \phi(\mathbf{B})$ for $\mathbf{B} \equiv \{\mathbf{B}(t; \eta, \sigma^2, x) : t \geq 0\}$, where ϕ is the one-dimensional reflection map in (2.5) and $\mathbf{B}$ is Brownian motion. There is a relatively simple expression for the transient distribution of RBM when there is only a single barrier; see p. 49 of Harrison (1985).

Theorem 5.7.7. (transition probability of RBM with one reflecting barrier) *If $\mathbf{R} \equiv \{\mathbf{R}(t; \eta, \sigma^2, x) : t \geq 0\}$ is an (η, σ^2) -RBM then*

$$P(\mathbf{R}(t) \leq y | \mathbf{R}(0) = x) = 1 - \Phi\left(\frac{-y + x + \eta t}{\sigma\sqrt{t}}\right) - \exp(2\eta y / \sigma^2) \Phi\left(\frac{-y - x - \eta t}{\sigma\sqrt{t}}\right),$$

where Φ is the standard normal cdf.

We now observe that we can express RBM with negative drift (and one reflecting barrier at the origin) in terms of *canonical RBM* with drift coefficient -1 and diffusion coefficient 1 . We first state the result for Brownian motion and then for reflected Brownian motion.

Theorem 5.7.8. (scaling to canonical Brownian motion) *If $m < 0$ and $\sigma^2 > 0$, then*

$$\{a\mathbf{B}(bt; m, \sigma^2, x) : t \geq 0\} \stackrel{d}{=} \{\mathbf{B}(t; -1, 1, ax) : t \geq 0\} \quad (7.36)$$

and

$$\{\mathbf{B}(t; m, \sigma^2, x) : t \geq 0\} \stackrel{d}{=} \{a^{-1}\mathbf{B}(b^{-1}t; -1, 1, ax) : t \geq 0\} \quad (7.37)$$

for

$$\begin{aligned} a &= \frac{|m|}{\sigma^2} > 0, & b &= \frac{\sigma^2}{m^2} > 0, \\ m &= -\frac{1}{ab} < 0, & \sigma^2 &= \frac{1}{a^2b} > 0. \end{aligned} \quad (7.38)$$

Theorem 5.7.9. (scaling to canonical RBM). *If $\eta < 0$ and $\sigma^2 > 0$, then*

$$\{a\mathbf{R}(bt; \eta, \sigma^2, Y) : t \geq 0\} \stackrel{d}{=} \{\mathbf{R}(t; -1, 1, aY) : t \geq 0\} \quad (7.39)$$

and

$$\{\mathbf{R}(t; \eta, \sigma^2, Y) : t \geq 0\} \stackrel{d}{=} \{a^{-1}\mathbf{R}(b^{-1}t; -1, 1, aY) : t \geq 0\} \quad (7.40)$$

for

$$\begin{aligned} a &\equiv \frac{|\eta|}{\sigma^2} > 0, & b &\equiv \frac{\sigma^2}{\eta^2}, \\ \eta &= \frac{-1}{ab}, & \sigma^2 &= \frac{1}{a^2b}, \end{aligned} \quad (7.41)$$

as in (7.38) of Chapter 4.

Theorem 5.7.9 is significant because it implies that we only need to do calculations for a single RBM — canonical RBM. Expressions for the moments of canonical RBM are given Abate and Whitt (1987a,b) along with various approximations. There it is shown that the time-dependent moments can be characterized via cdf's. In particular, the time-dependent moments starting at 0, normalized by dividing by the steady-state moments are cdf's. Moreover the differences $E(\mathbf{R}(t)|\mathbf{R}(0) = x) - E[\mathbf{R}(t)|\mathbf{R}(0) = 0]$ divided by x are complementary cdf's (ccdf's), and all these cdf's have revealing structure. Here are explicit expressions for the first two moments.

Theorem 5.7.10. (moments of canonical RBM) *If $\mathbf{R}$ is canonical RBM, then*

$$\begin{aligned} E[\mathbf{R}(t)|\mathbf{R}(0) = x] &= 2^{-1} + \sqrt{t}\phi\left(\frac{t-x}{\sqrt{t}}\right) \\ &\quad - (t-x+2^{-1})\left[1 - \Phi\left(\frac{t-x}{\sqrt{t}}\right)\right] \\ &\quad - 2^{-1}e^{2x}\left[1 - \Phi\left(\frac{t+x}{\sqrt{t}}\right)\right] \end{aligned}$$

and

$$\begin{aligned} E[\mathbf{R}(t)^2|\mathbf{R}(0) = x] &= 2^{-1} + ((x-1)\sqrt{t} - \sqrt{t^3})\phi\left(\frac{t-x}{\sqrt{t}}\right) \\ &\quad + ((t-x)^2 + t - 2^{-1})\left[1 - \Phi\left(\frac{t-x}{\sqrt{t}}\right)\right] \\ &\quad + e^{2x}(t+x-2^{-1})\left[1 - \Phi\left(\frac{t+x}{\sqrt{t}}\right)\right], \end{aligned}$$

where Φ and ϕ are the standard normal cdf and pdf.

When thinking about RBM approximations for queues, it is sometimes useful to regard RBM as a special M/M/1 queue with $\rho = 1$. After doing appropriate scaling, the M/M/1 queue-length process approaches a nondegenerate limit as $\rho \rightarrow 1$. Thus structure of RBM can be deduced from structure for the M/M/1 queue; see Abate and Whitt (1988a-d). This is one way to characterize the covariance function of stationary RBM; see Abate and Whitt (1988c). Recall that a nonnegative-real-valued function f is completely monotone if it has derivatives of all orders that alternate in sign. Equivalently, f can be expressed as a mixture of exponential distributions; see p. 439 of Feller (1971).

Theorem 5.7.11. (covariance function of RBM) *Let $\mathbf{R}^*$ be canonical RBM initialized by giving $\mathbf{R}^*(0)$ an exponential distribution with mean $1/2$. The process $\mathbf{R}^*$ is a stationary process with completely monotone covariance function*

$$\begin{aligned} \text{Cov}(\mathbf{R}^*(0), \mathbf{R}^*(t)) &\equiv E[\mathbf{R}^*(t) - 2^{-1}](\mathbf{R}^*(0) - 2^{-1})] \\ &= 2(1 - 2t - t^2)[1 - \Phi(\sqrt{t})] + 2\sqrt{t}(1 + t)\phi(\sqrt{t}) \\ &= H_{1e}^c(t) = H_2^c(t), \quad t \geq 0, \end{aligned}$$

where H_k is the k^{th} -moment cdf and H_{1e}^c is the stationary-excess cdf associated with the first-moment cdf, i.e.,

$$H_k(t) \equiv \frac{E[\mathbf{R}(t)^k | \mathbf{R}(0) = 0]}{E\mathbf{R}(\infty)^k}, \quad t \geq 0,$$

and

$$H_{1e}^c(t) \equiv 1 - 2 \int_0^t H_1^c(s) ds, \quad t \geq 0.$$

Canonical RBM has asymptotic variance

$$\sigma_{\mathbf{R}}^2 \equiv \lim_{t \rightarrow \infty} t^{-1} \text{Var} \left(\int_0^t \mathbf{R}(s) ds | \mathbf{R}(0) = x \right) = 1/2.$$

5.7.5. First-Passage Times

We can also establish limits for first passage times. For a stochastic process $\{Z(t) : t \geq 0\}$, let $T_{a,b}(Z)$ denote the first passage time for Z to go from a to b . (We assume that $Z(0) = a$, and consider the first passage time to b .) In general, the first passage time functional is not continuous on D or even on the subset C , but the first passage time functional is continuous

almost surely with respect to BM or RBM, because BM and RBM cross any level w.p.1 in a neighborhood of any time that they first hit a level. Hence we can invoke a version of the continuous mapping theorem to conclude that limits holds for the first passage times.

Theorem 5.7.12. (limits for first passage times) *Under the assumptions of Theorem 5.7.1,*

$$\frac{T_{a\sqrt{n}, b\sqrt{n}}(W_n)}{n} \Rightarrow T_{a,b}(\mathbf{W})$$

for any positive a, b with $a \neq b$ and $0 \leq a, b \leq K$, where $\mathbf{W}$ is RBM and W_n is the unnormalized workload process in model n .

Now let $T_{a,b}(\mathbf{R})$ be the first-passage time from a to b for one-sided canonical RBM. The first-passage time upward is the same as when there is a (higher) upper barrier (characterized in Theorems 5.7.5 and 5.7.6), but the first-passage time down is new. Let $f(t; a, b)$ be the pdf of $T_{a,b}(\mathbf{R})$ and let $\hat{f}(s; a, b)$ be its Laplace transform, i.e.,

$$\hat{f}(s; a, b) \equiv \int_0^\infty e^{-st} f(t; a, b) dt ,$$

where s is a complex variable with positive real part. The Laplace transforms to and from the origin have a relatively simple form; see Abate and Whitt (1988a). Again, numerical transform inversion can be applied to compute the probability distributions themselves.

Theorem 5.7.13. (RBM first-passage-time transforms and moments) *For canonical RBM (with no upper barrier), the first-passage-time Laplace transforms to and from the origin are, respectively,*

$$\hat{f}(s; x, 0) = e^{-xr_2}$$

and

$$\hat{f}(s; 0, x) = \frac{r_1 + r_2}{r_1 e^{-xr_2} + r_2 e^{xr_1}}$$

for

$$r_1(s) = 1 + \sqrt{1 + 2s} \quad \text{and} \quad r_2(s) = \sqrt{1 + 2s} - 1 ,$$

so that

$$\begin{aligned} ET_{x,0} &= x, & Var T_{x,0} &= x , \\ ET_{0,x} &= 2^{-1}[e^{2x} - 1 - 2x] & \text{and} \\ Var T_{0,x} &= 4^{-1}[e^{4x} - 1 - 4x + 4e^{2x}(1 - 2x) - 4] . \end{aligned}$$

The first passage time down is closely related to the busy period of a queue, i.e., the time from when a buffer first becomes nonempty until it becomes empty again. This concept is somewhat more complicated for fluid queues than standard queues. In either case, the distribution of the busy period for small values tends to depend on the fine structure of the model, but the tail of the busy period often can be approximated robustly, and Brownian approximations can play a useful role; see Abate and Whitt (1988d, 1995b).

First-passage-time cdf's are closely related to extreme-value cdf's because $T_{0,a}(W) \leq t$ if and only if $W^\uparrow(t) \equiv \sup_{0 \leq s \leq t} W(s) \geq a$. Extreme-value theory shows that there is statistical regularity associated with both first-passage times and extreme values as $t \rightarrow \infty$ and $a \rightarrow \infty$; see Resnick (1987). Heavy-traffic extreme-value approximations for queues are discussed by Berger and Whitt (1995a), Glynn and Whitt (1995) and Chang (1997). A key limit is

$$2R^\uparrow(t) - \log(2t) \Rightarrow Z \quad \text{as } t \rightarrow \infty ,$$

where R is canonical RBM and Z has the *Gumbel cdf*, i.e.,

$$P(Z \leq x) \equiv \exp(-e^{-x}), \quad -\infty < x < \infty .$$

This limit can serve as a basis for extreme-value engineering.

To summarize, in this section we have displayed Brownian limits for a fluid queue, obtained by combining the general fluid-queue limits in Theorem 5.4.1 with the multidimensional version of Donsker's theorem in Theorem 4.3.5. We have also displayed various formulas for RBM that are helpful in applications of the Brownian limit. We discuss RBM limits and approximations further in the next section and in Sections 8.4 and 9.6.

5.8. Planning Queueing Simulations

In this section, following Whitt (1989a), we see how the Brownian approximation stemming from the Brownian heavy-traffic limit in Section 5.7 can be applied to plan simulations of queueing models. In particular, we show how the Brownian approximation can be used to estimate the required simulation run lengths needed to obtain desired statistical precision, before any data have been collected. These estimates can be used to help design the simulation experiment and even to determine whether or not a contemplated experiment should be conducted.

The queueing simulations considered are single replications (one long run) of a single queue conducted to estimate steady-state characteristics,

such as long-run-average steady-state workload. For such simulations to be of genuine interest, the queueing model should be relatively complicated, so that exact numerical solution is difficult. On the other hand, the queueing model should be sufficiently tractable that we can determine an appropriate Brownian approximation.

We assume that both these criteria are met. Indeed, we specify the models that we consider by stipulating that scaled versions of the stochastic process of interest, with the standard normalization, converge to RBM as $\rho \uparrow 1$. For simplicity, we focus on the workload process in a fluid queue with infinite capacity, but the approach applies to other models as well.

Of course, such a Brownian approximation directly yields an approximation for the steady-state performance, but nevertheless we may be interested in the additional simulation in order to develop a more precise understanding of the steady-state behavior. Indeed, one use of such simulations is to evaluate how various candidate approximations perform. Then we often need to perform a large number of simulations in order to see how the approximations perform over a range of possible model parameters.

In order to exploit the Brownian approximation for a single queue, we focus on simulations of a single queue. However, the simulation actually might be for a network of queues. Then the analysis of a single queue is intended to apply to any one queue in that network. If we want to estimate the steady-state performance at all queues in the network, then the required simulation run length for the network would be the maximum required for any one queue in the network. Our analysis shows that it often suffices to focus on the bottleneck (most heavily loaded) queue in the network.

At first glance, the experimental design problem may not seem very difficult. To get a rough idea about how long the runs should be, one might do one “pilot” run to estimate the required simulation run lengths. However, such a preliminary experiment requires that you set up the entire simulation before you decide whether or not to conduct the experiment. Nevertheless, if such a sampling procedure could be employed, then the experimental design problem would indeed not be especially difficult. Interest stems from the fact that one sample run can be extremely misleading.

This queueing experimental design problem is interesting and important primarily because a uniform allocation of data over all cases (parameter values) is not nearly appropriate. Experience indicates that, for given statistical precision, the required amount of data increases as the traffic intensity increases and as the arrival-and-service variability (appropriately quantified) increases. Our goal is to quantify these phenomena.

To quantify these phenomena, we apply the space and time scaling func-

tions. Our analysis indicates that to achieve a uniform relative error over all values of the traffic intensity ρ that the run length should be approximately proportional to the time-scaling factor $(1 - \rho)^{-2}$ (for sufficiently high ρ). Relative error appears to be a good practical measure of statistical precision, except possibly when very small numbers are involved. Then absolute error might be preferred. It is interesting that the required run length depends strongly on the criterion used. With the absolute error criterion, the run length should be approximately proportional to $(1 - \rho)^{-4}$. With either the relative or absolute error criteria, there obviously are great differences between the required run lengths for different values of ρ , e.g., for $\rho = 0.8$, 0.9 and 0.99.

We divide the simulation run-length problem into two components. First, there is the question: What should be the required run length given that the system starts in equilibrium (steady state)? Second, there is the question: What should we do in the customary situation in which it is not possible to start in equilibrium? We propose to delete an initial portion of each simulation run before collecting data in order to allow the system to (approximately) reach steady state. By that method, we reduce the bias (the systematic error that occurs when the expected value of the estimator differs from the quantity being estimated). The second question, then, can be restated as: How long should be the initial segment of the simulation run that is deleted?

Focusing on the first question first, we work with the workload stochastic process, assuming that we have a stationary version, denoted by W_ρ^* . First, however, note that specifying the run length has no meaning until we specify the time units. To fix the time units, we assume that the output rate in the queueing system is μ . (It usually suffices to let $\mu = 1$, but we keep general μ to show how it enters in.)

For the general fluid-queue model we have the RBM approximation in (7.8). However, since we are assuming that we start in equilibrium, instead of the Brownian approximation in (7.8), we assume that we have the associated *stationary Brownian approximation*

$$\{W_\rho^*(t) : t \geq 0\} \approx \{\sigma_X^2 \mu^{-1} (1 - \rho)^{-1} \mathbf{R}^*(\sigma_X^{-2} \mu^2 (1 - \rho)^2 t; -1, 1) : t \geq 0\}, \quad (8.1)$$

where σ_X^2 is the variability parameter, just as in (7.8), and $\mathbf{R}^*$ is a stationary version of canonical RBM, with initial exponential distribution, i.e.,

$$\{\mathbf{R}^*(t; -1, 1) : t \geq 0\} \stackrel{d}{=} \{\mathbf{R}(t; -1, 1, Y) : t \geq 0\}, \quad (8.2)$$

where the initial position Y is an exponential random variable with mean

1/2 independent of the standard Brownian motion being reflected; i.e., $\mathbf{R}^* = \phi(\mathbf{B} + Y)$ where ϕ is the reflection map and $\mathbf{B}$ is a standard Brownian motion independent of the exponential random variable Y .

The obvious application is with $\{W_\rho^*(t) : t \geq 0\}$ being a stationary version of a workload process, as defined in Section 5.2. However, our analysis applies to any stationary process having the Brownian approximation in (8.1).

5.8.1. The Standard Statistical Procedure

To describe the standard statistical procedure, let $\{W(t) : t \geq 0\}$ be a stochastic process of interest and assume that is stationary with $EW(t)^2 < \infty$. (We use that notation because we are thinking of the workload process, but the statistical procedure is more general, not even depending upon the Brownian approximation.) Our object is to estimate the mean $E[W(0)]$ by the *sample mean*, i.e., by the time average

$$\bar{W}_t \equiv t^{-1} \int_0^t W(s) ds, \quad t \geq 0. \quad (8.3)$$

The standard statistical procedure, assuming ample data, is based on a CLT for $\bar{W}_t$. We assume that

$$t^{1/2}(\bar{W}_t - E[W(0)]) \Rightarrow N(0, \sigma^2) \quad \text{as } t \rightarrow \infty, \quad (8.4)$$

where σ^2 is the *asymptotic variance*, defined by

$$\sigma^2 \equiv \lim_{t \rightarrow \infty} t \text{Var}(\bar{W}_t) = 2 \int_0^\infty C(t) dt, \quad (8.5)$$

and $C(t)$ is the (auto) *covariance function*

$$C(t) \equiv E[W(t)W(0)] - (E[W(0)])^2, \quad t \geq 0. \quad (8.6)$$

Of course, a key part of assumption (8.4) is the requirement that the asymptotic variance σ^2 be finite. The CLT in (8.4) is naturally associated with a Brownian approximation for the process $\{W(t) : t \geq 0\}$. Such CLTs for stationary processes with weak dependence were discussed in Section 4.4. Based on (8.4), we use the normal approximation

$$\bar{W}_t \approx N(E[W(0)], \sigma^2/t) \quad (8.7)$$

for the (large) t of interest, where σ^2 is the asymptotic variance in (8.5).

Based on (8.7), a $[(1-\beta) \cdot 100]\%$ *confidence interval* for the mean $E[W(0)]$ is

$$[\bar{W}_t - z_{\beta/2}(\sigma^2/t)^{1/2}, \bar{W}_t + z_{\beta/2}(\sigma^2/t)^{1/2}] , \quad (8.8)$$

where

$$P(-z_{\beta/2} \leq N(0, 1) \leq z_{\beta/2}) = 1 - \beta . \quad (8.9)$$

The width of the confidence interval in (8.8) provides a natural measure of the *statistical precision*. There are two natural criteria to consider: *absolute width* and *relative width*. Relative width looks at the ratio of the width to the quantity to be estimated, $E[W(0)]$.

For any given β , the absolute width and relative width of the $[(1-\beta) \cdot 100]\%$ confidence intervals for the mean $E[W(0)]$ are, respectively,

$$w_a(\beta) = \frac{2\sigma z_{\beta/2}}{t^{1/2}} \quad \text{and} \quad w_r(\beta) = \frac{2\sigma z_{\beta/2}}{t^{1/2} E[W(0)]} . \quad (8.10)$$

For specified *absolute width* ϵ and specified *confidence level* $1-\beta$, the required simulation run length, given (8.7), is

$$t_a(\epsilon, \beta) = \frac{4\sigma^2 z_{\beta/2}^2}{\epsilon^2} . \quad (8.11)$$

For specified *relative width* ϵ and specified *confidence level* $1-\beta$, the required length of the estimation interval, given (8.7), is

$$t_r(\epsilon, \beta) = \frac{4\sigma^2 z_{\beta/2}^2}{\epsilon^2 (E[W(0)])^2} . \quad (8.12)$$

From (8.11) and (8.12) we draw the important and well-known conclusion that both $t_a(\epsilon, \beta)$ and $t_r(\epsilon, \beta)$ are inversely proportional to ϵ^2 and directly proportional to σ^2 and $z_{\beta/2}^2$.

Standard statistical theory describes how observations can be used to estimate the unknown quantities $E[W(0)]$ and σ^2 . Instead, we apply additional information about the model to obtain rough preliminary estimates for $E[W(0)]$ and σ^2 without data.

5.8.2. Invoking the Brownian Approximation

At this point we invoke the Brownian approximation in (8.1). We assume that the process of interest is W_ρ^* and that it can be approximated by scaled stationary canonical RBM as in (8.1). The steady-state mean of canonical

RBM and its asymptotic variance are both $1/2$; see Theorems 5.7.10 and 5.7.11. It thus remains to consider the scaling.

To consider the effect of scaling space and time in general, let W again be a general stationary process with covariance function C and let

$$W_{y,z}(t) \equiv yW(zt), \quad t \geq 0$$

for $y, z > 0$. Then the mean $E[W_{y,z}(t)]$, covariance function $C_{y,z}(t)$ and asymptotic variance of $W_{y,z}$ are, respectively,

$$\begin{aligned} E[W_{y,z}(t)] &= yEW(zt) = yE[W(t)] , \\ C_{y,z}(t) &= y^2C(zt) \quad \text{and} \quad \sigma_{y,z}^2 = y^2\sigma^2/z . \end{aligned} \quad (8.13)$$

Thus, from (8.1) and (8.13), we obtain the important approximations

$$E[W_\rho^*(0)] \approx \frac{\sigma_X^2}{2\mu(1-\rho)} \quad \text{and} \quad \sigma_{W_\rho^*}^2 \approx \frac{\sigma_X^6}{2\mu^4(1-\rho)^4} . \quad (8.14)$$

We have compared the approximation for the mean in (8.14) to the exact formula for the M/G/1 workload process in Example 5.7.1. Similarly, the exact formula for the asymptotic variance for the M/M/1 workload process, where $\mu = 1$, is

$$\sigma_{W_\rho}^2 = \frac{2\rho(3-\rho)}{(1-\rho)^4} ; \quad (8.15)$$

see (23) of Whitt (1989a). Formula (8.15) reveals limitations of the approximation in (8.14) in light traffic (as $\rho \downarrow 0$), but formula (8.15) agrees with the approximation in (8.14) in the limit as $\rho \rightarrow 1$, because $\sigma_X^2 = 2\rho$ for the M/M/1 queue; let $EV_1 = 1$ and $c_V^2 = 1$ in (7.20). Numerical comparisons of the predictions with simulation estimates in more general models appear in Whitt (1989a). These formulas show that the approximations give good rough approximations for ρ not too small (e.g., for $\rho \geq 1/2$).

Combining (8.12) and (8.14), we see that the approximate required simulation run length for W_ρ^* given a specified *relative* width ϵ and confidence level $1 - \beta$ for the confidence interval for $E[W_\rho^*(0)]$ is

$$t_r(\epsilon, \beta) \approx \frac{8\sigma_X^2 z_{\beta/2}^2}{\epsilon^2 \mu^2 (1-\rho)^2} . \quad (8.16)$$

Combining (8.11) and (8.14), we see that the approximate required simulation run length for W_ρ^* given a specified *absolute* width ϵ and confidence

level $1 - \beta$ for the confidence interval for $E[W_\rho(0)]$ is

$$t_a(\epsilon, \beta) \approx \frac{2\sigma_X^6 z_{\beta/2}^2}{\epsilon^2 \mu^4 (1 - \rho)^4} . \quad (8.17)$$

In summary, the Brownian approximation in (8.1) dictates that, with a criterion based on the relative width of the confidence interval, the required run length should be directly proportional to both the time-scaling term as a function of ρ alone, $(1 - \rho)^{-2}$, and the heavy-traffic variability parameter σ_X^2 . In contrast, with the absolute standard error criterion, the required run length should be directly proportional to $(1 - \rho)^{-4}$, the *square* of the time-scaling term as a function of ρ alone, and σ_X^6 , the *cube* of the heavy-traffic variability parameter σ_X^2 .

The second question mentioned at the outset is: How to determine an initial transient portion of the simulation run to delete? To develop an approximate answer, we can again apply the Brownian approximation in (8.1). If the system starts empty, we can consider canonical RBM starting empty. By Theorem 5.7.10, the time-dependent mean of canonical RBM $E[\mathbf{R}(t) | \mathbf{R}(0) = 0]$ is within about 1% of its steady-state mean $1/2$ at $t = 4$. Hence, if we were simulating canonical RBM, then we might delete an initial portion of length 4. Thus, by (8.1), a rough rule of thumb for the queueing process W_ρ (with unit processing rate) is to delete an initial segment of length $4\sigma_X^2/\mu^2(1 - \rho)^2$. When we compare this to formula (8.16), we see that the proportion of the total run that should be deleted should be about $\epsilon^2/2z_{\beta/2}^2$, which is small when ϵ is small.

We can also employ the Brownian approximation to estimate the bias due to starting away from steady-state. For example, the bias due to starting empty with canonical RBM is

$$\begin{aligned} E\bar{\mathbf{R}}_t - 1/2 &= t^{-1} \int_0^t (E[\mathbf{R}(s); -1, 1, 0] - 1/2) ds \\ &\approx t^{-1} \int_0^\infty (E[\mathbf{R}(s); -1, 1, 0] - 1/2) ds = 1/4t , \end{aligned} \quad (8.18)$$

by Corollary 1.3.4 of Abate and Whitt (1987a). The approximate relative bias is thus $1/2t$. That same relative bias should apply approximately to the workload process in the queue. We can also estimate the reduced bias due to deleting an initial portion of the run, using Theorem 5.7.10 and the hyperexponential approximation

$$1/2 - E[\mathbf{R}(t; -1, 1, 0)] \approx 0.36e^{-5.23t} + 0.138e^{-0.764t}, \quad t \geq 0 . \quad (8.19)$$

Our entire analysis depends on the normal approximation in (8.7), which in turn depends on the simulation run length t . Not only must t be sufficiently large so that the estimated statistical precision based on (8.7) is adequate, but t must be sufficiently large so that the normal approximation in (8.7) is itself reasonable. Consistent with intuition, experience indicates that the run length required for (8.7) to be a reasonable approximation also depends on the parameters ρ and σ_X^2 , with t needing to increase as ρ and σ_X^2 increase. We can again apply the Brownian approximation to estimate the run length required. We can ask what run length is appropriate for a normal approximation to the distribution of the sample mean of canonical RBM. First, however, the time scaling alone tells us that the run length must be at least of order $\sigma_X^2/\mu^2(1-\rho)^2$. This rough analysis indicates that the requirement for (8.7) to be a reasonable approximation is approximately the same as the requirement to control the relative standard error. For further analysis supporting this conclusion, see Asmussen (1992).

5.9. Heavy-Traffic Limits for Other Processes

We now obtain heavy-traffic stochastic-process limits for other processes besides the workload process in the setting of Section 5.4. Specifically, we obtain limits for the departure process and the processing time.

5.9.1. The Departure Process

We first obtain limits for the departure process defined in (2.11), but in general we can have difficulties applying the continuous-mapping approach with addition starting from (2.11) because the limit processes $\mathbf{S}$ and $-\mathbf{L}$ can have common discontinuities of opposite sign. We can obtain positive results when we rule that out, again invoking Theorem 12.7.3.

Let the scaled departures processes be defined by

$$\mathbf{D}_n \equiv c_n^{-1}(D_n(nt) - \mu_n nt), \quad t \geq 0. \quad (9.1)$$

Theorem 5.9.1. (limit for the departure process) *Let the conditions of Theorem 5.4.1 hold. If the topology on D is J_1 , assume that $\mathbf{S}$ and $\mathbf{L}$ almost surely have no common discontinuities. If the topology on D is M_1 , assume that $\mathbf{S}$ and $\mathbf{L}$ almost surely have no common discontinuities with jumps of*

common sign. Then, jointly with the limits in (4.5) and (4.7),

$$\mathbf{D}_n \Rightarrow \mathbf{D} \equiv \mathbf{S} - \mathbf{L} \quad \text{in } D \quad (9.2)$$

with the same topology, for $\mathbf{D}_n$ in (9.1), $\mathbf{S}$ in (4.5) and $\mathbf{L}$ in (4.9).

Proof. By (2.11),

$$\mathbf{D}_n = \mathbf{S}_n - \mathbf{L}_n .$$

By Theorem 5.4.1, $(\mathbf{S}_n, \mathbf{L}_n) \Rightarrow (\mathbf{S}, \mathbf{L})$ in D^2 jointly with the other limits. Just as in the proof of Theorem 5.4.1, we can apply the continuous mapping theorem, Theorem 3.4.3, with addition. Under the conditions on the discontinuities of $\mathbf{S}$ and $\mathbf{L}$, addition is measurable and almost surely continuous. Hence we obtain the desired limit in (9.2). ■

The extra assumption in Theorem 5.9.1 is satisfied when $P(S_n(t) = \mu_n t, \quad t \geq 0) = 1$ or when $\mathbf{X}$ has no negative jumps (which implies that $\mathbf{L} \equiv \psi_L(\mathbf{X})$ has continuous paths).

As an alternative to (9.1), we can use the input rate λ_n in the translation term of the normalized departure process; i.e., let

$$\mathbf{D}'_n \equiv c_n^{-1}(D_n(nt) - \lambda_n nt), \quad t \geq 0 . \quad (9.3)$$

When the input rate appears in the translation term, we can directly compare the departure processes D_n to the cumulative-input processes C_n .

Corollary 5.9.1. (limit for the departure process with input centering) *Under the assumptions of Theorem 5.9.1,*

$$\mathbf{D}'_n \Rightarrow \mathbf{D}' \equiv -\eta \mathbf{e} + \mathbf{S} - \mathbf{L} \quad \text{in } (D, M_1) \quad (9.4)$$

for $\mathbf{D}'_n$ in (9.3), η in (4.6), $\mathbf{e}(t) \equiv t$ for $t \geq 0$, $\mathbf{S}$ in (4.5) and $\mathbf{L}$ in (4.9).

Proof. Note that $\mathbf{D}'_n = \mathbf{D}_n - \eta_n \mathbf{e}$. Hence, as before, we can apply the continuous-mapping theorem, Theorem 3.4.3, with addition to the joint limit $(\mathbf{D}_n, \eta_n \mathbf{e}) \Rightarrow (\mathbf{D}, \eta \mathbf{e})$, which holds by virtue of Theorems 5.9.1 and 11.4.5. ■

5.9.2. The Processing Time

We now establish heavy-traffic limits for the processing time $T(t)$ in (2.12). We first exploit (2.13) when $K = \infty$. Let the scaled processing-time processes be

$$\mathbf{T}_n(t) \equiv c_n^{-1} T_n(nt), \quad t \geq 0 . \quad (9.5)$$

Theorem 5.9.2. (limit for the processing time when $K = \infty$) *Suppose that, in addition to the conditions of Theorem 5.4.1, $K = \infty$, $\mu_n \rightarrow \mu$ as $n \rightarrow \infty$, where $0 < \mu < \infty$,*

$$\eta_{C,n} \equiv n(\lambda_n - \mu)/c_n \rightarrow \eta_C \quad (9.6)$$

and

$$\eta_{S,n} \equiv n(\mu_n - \mu)/c_n \rightarrow \eta_S, \quad (9.7)$$

where $-\infty < \eta_C < \infty$ and $-\infty < \eta_S < \infty$, so that $\eta = \eta_C - \eta_S$. If the topology on D is J_1 , suppose that almost surely no two of the limit processes $\mathbf{C}$, $\mathbf{S}$ and $\mathbf{L}$ have common discontinuities. If the topology on D is M_1 , assume that $\mathbf{L}$ and $\mathbf{C}$ almost surely have no common discontinuities with jumps of opposite sign, and $\mathbf{S}$ and $\mathbf{L}$ almost surely have no common discontinuities with jumps of common sign. Suppose that

$$P(\mathbf{S}(0) = 0) = 1. \quad (9.8)$$

Then

$$\mathbf{T}_n \Rightarrow \mu^{-1} \mathbf{W} \quad \text{in } D \quad (9.9)$$

with the same topology on D , jointly with the limits in (4.5) and (4.7), for $\mathbf{T}_n$ in (9.5) and $\mathbf{W}$ in (4.9) with $K = \infty$.

Proof. We can apply the continuous-mapping approach with first passage times, using the inverse map with centering in Section 13.7. Specifically, we can apply Theorem 13.7.4 with the Skorohod representation theorem, Theorem 3.2.2. From (2.13),

$$n^{-1}T_n(nt) + nt = \inf\{u \geq 0 : n^{-1}S_n(nu) \geq n^{-1}(C_n(nt) + W'_n(0) + L_n(nt))\}.$$

By (4.5), (9.6) and (9.7),

$$(n/c_n)(\hat{\mathbf{S}}_n - \mu \mathbf{e}, \hat{\mathbf{Z}}_n - \mu \mathbf{e}) \Rightarrow (\mathbf{S} + \eta_S \mathbf{e}, \mathbf{Z} + \eta_C \mathbf{e}), \quad (9.10)$$

where

$$\hat{\mathbf{S}}_n \equiv n^{-1}S_n(nt) \quad \text{and} \quad \hat{\mathbf{Z}}_n \equiv n^{-1}(C_n(nt) + W'_n(0) + L_n(nt)), \quad t \geq 0.$$

We use the conditions on the discontinuities of $\mathbf{C}$ and $\mathbf{L}$ to obtain the limit

$$(n/c_n)(\hat{\mathbf{Z}}_n - \mu \mathbf{e}) \Rightarrow \mathbf{Z} + \eta_C \mathbf{e},$$

where

$$\mathbf{Z} \equiv \mathbf{C} + W'(0) + \mathbf{L},$$

by virtue of Theorem 12.7.3. Since

$$\hat{\mathbf{T}}_n(t) \equiv n^{-1}T_n(nt) = (\hat{\mathbf{S}}_n^{-1} \circ \hat{\mathbf{Z}}_n)(t) - t, \quad t \geq 0, \quad (9.11)$$

the desired limit for $\mathbf{T}_n$ follows from Theorem 13.7.4. In particular, (9.10), (9.11) and (9.8) imply the limit

$$\begin{aligned} & (n/c_n)(\hat{\mathbf{S}}_n^{-1} \circ \hat{\mathbf{Z}}_n - \mu^{-1}\mathbf{e} \circ \mu\mathbf{e}) \\ & \Rightarrow \frac{(\mathbf{Z} + \eta_C\mathbf{e}) - (\mathbf{S} + \eta_S\mathbf{e}) \circ \mu^{-1}\mathbf{e} \circ \mu\mathbf{e}}{\mu} = \frac{\mathbf{W}}{\mu}. \quad \blacksquare \end{aligned}$$

The continuity conditions in Theorem 5.9.2 are satisfied when $\mathbf{S}$ is almost surely continuous and $\mathbf{X}$ almost surely has no negative jumps (which makes $\mathbf{L}$ almost surely have continuous paths). That important case appears in the convergence to reflected stable Lévy motion in Theorem 8.5.1.

We can also obtain a FCLT for T_n when $K < \infty$ under stronger continuity conditions and pointwise convergence under weaker conditions. (It may be possible to establish analogs to part (b) below without such strong continuity conditions.)

Theorem 5.9.3. (limits for the processing time when $K \leq \infty$) *Suppose that the conditions of Theorem 5.4.1 hold with $0 < K \leq \infty$ and $\mu_n \rightarrow \mu$, where $0 < \mu < \infty$.*

(a) *If*

$$P(t \in \text{Disc}(\mathbf{S})) = P(t \in \text{Disc}(\mathbf{W})) = 0, \quad (9.12)$$

then

$$\mathbf{T}_n(t) \Rightarrow \mu^{-1}\mathbf{W}(t) \quad \text{in } \mathbb{R}. \quad (9.13)$$

(b) *If*

$$P(\mathbf{C} \in C) = P(\mathbf{S} \in C) = 1, \quad (9.14)$$

then

$$\mathbf{T}_n \Rightarrow \mu^{-1}\mathbf{W} \quad \text{in } (D, M_1), \quad (9.15)$$

where $P(\mathbf{W} \in C) = 1$.

Proof. (a) By (2.12),

$$n^{-1}T_n(nt) = \inf\{u \geq 0 : S_n(n(t+u)) - S_n(nt) \geq W_n(nt)\},$$

so that

$$\mathbf{T}_n(t) = \inf\{u \geq 0 : \mu_n u + \mathbf{S}_n(t + u(c_n/n)) - \mathbf{S}_n(t) \geq \mathbf{W}_n(t)\}.$$

By the continuous-mapping approach, with condition (9.12),

$$\mathbf{T}_n(t) \Rightarrow \inf\{u \geq 0 : \mu u \geq \mathbf{W}(t)\},$$

which implies the conclusion in (9.13).

(b) Under condition (9.14),

$$\sup_{0 \leq t \leq T} \{|\mathbf{S}_n(t + u(c_n/n)) - \mathbf{S}_n(t)|\} \rightarrow 0 \quad \text{as } n \rightarrow \infty \quad \text{w.p.1}$$

for any T with $0 < T < \infty$; see Section 12.4. Hence the conclusion in part (a) holds uniformly over all bounded intervals. An alternative proof follows the proof of Theorem 5.9.2, including the process $\{U(t) : t \geq 0\}$ when $K < \infty$. ■

Remark 5.9.1. *The heavy-traffic snapshot principle.* With the previous heavy-traffic theorems in this section, Theorems 5.9.2 and 5.9.3 establish a version of the heavy-traffic snapshot principle, a term coined by Reiman (1982): *In the heavy-traffic limit, the processing time is asymptotically negligible compared to the time required for the workloads to change significantly.* Since time is scaled by n , the workloads can change significantly only over time intervals of length of order n . On the other hand, since the space scaling is by c_n , where $c_n \rightarrow \infty$ but $c_n/n \rightarrow 0$ as $n \rightarrow \infty$, the workload itself tends to be only of order c_n , which is asymptotically negligible compared to n . Correspondingly, Theorems 5.9.2 and 5.9.3 show that that processing times also are of order c_n . Thus, in the heavy-traffic limit, the workload when a particle of work departs is approximately the same as the workload when that particle of work arrived.

The heavy-traffic snapshot principle also holds in queueing networks. Thus the workload seen upon each visit to a queue in the network and upon departure from the network by a particle flowing through the network is the same, in the heavy-traffic limit, as seen by that particle upon initial arrival. The heavy-traffic snapshot principle implies that network status can be communicated effectively in a heavily loaded communication network: A special packet sent from source to destination may record the buffer content at each queue on its path. Then this information may be passed back to the source by a return packet. The snapshot principle implies that the buffer contents at the queues will tend to remain near their original levels (relative to heavy-loading levels), so that the information does not become stale. (A caveat: With the fluid-limit scaling in Section 5.3, the heavy-traffic snapshot principle is not valid. In practice, we need to check if the snapshot principle applies.) For more on the impact of old information on scheduling service in queues, see Mitzenmacher (1997).

5.10. Priorities

In this book we primarily consider the standard first-come first-served (FCFS) service discipline in which input is served in order of arrival, but it can be important to consider other service disciplines to meet performance goals. We now illustrate how we can apply heavy-traffic stochastic-process limits to analyze a queue with a non-FCFS service discipline. Specifically, we now consider the fluid-queue model with priority classes. We consider the relatively tractable *preemptive-resume priority discipline*; i.e., higher-priority work immediately preempts lower-priority work and lower-priority work resumes service where it stopped when it regains access to the server. Heavy-traffic limits for the standard single-server queue with the preemptive-resume priority discipline were established by Whitt (1971a).

In general, there may be any number m of priority classes, but it suffices to consider only two because, from the perspective of any given priority class, all lower priority work can be ignored, and all higher-priority work can be lumped together. Thus, the model we consider now is the same as in Section 5.2 except that there are two priority classes. Let class 1 have priority over class 2. For $i = 1, 2$, there is a class- i cumulative-input stochastic process $\{C_i(t) : t \geq 0\}$. As before, there is a single server, a buffer with capacity K and a single service process $\{S(t) : t \geq 0\}$. (There is only a single shared buffer, not a separate buffer for each class.)

Like the polling service discipline considered in Section 2.4.2, the preemptive-resume priority service discipline is a work-conserving service policy. Thus the total workload process is the same as for the FCFS discipline considered above. We analyze the priority model to determine the performance enhancement experienced by the high-priority class and the performance degradation experienced by the low-priority class.

We first define class- i available-processing processes by letting

$$\begin{aligned} S_1(t) &\equiv S(t), \\ S_2(t) &\equiv S_1(t) - D_1(t), \end{aligned} \tag{10.1}$$

where $D_1 \equiv \{D_1(t) : t \geq 0\}$ is the class-1 departure process, defined as in (2.11). We then can define the class- i potential-workload processes by

$$X_i(t) \equiv W_i(0) + C_i(t) - S_i(t), \tag{10.2}$$

just as in (2.4). Then the class- i workload, overflow and departure processes are $W_i \equiv \phi_K(X_i)$, $U_i \equiv \psi_U(X_i)$ and $D_i \equiv S_i - \psi_L(X_i)$, just as in Section 5.2.

We now want to consider heavy-traffic limits for the two-priority fluid-queue model. As in Section 5.4, we consider a sequence of queues indexed by n . Suppose that the per-class input rates $\lambda_{1,n}$ and $\lambda_{2,n}$ and a maximum-potential output rate μ_n are well defined for each n , with limits as in (4.1) and (4.2). Then the class- i traffic intensity in model n is

$$\rho_{i,n} \equiv \lambda_{i,n}/\mu_n \quad (10.3)$$

and the overall traffic intensity in model n is

$$\rho_n \equiv \rho_{1,n} + \rho_{2,n} . \quad (10.4)$$

As a regularity condition, we suppose that $\mu_n \rightarrow \mu$ as $n \rightarrow \infty$, where $0 < \mu < \infty$.

In this context, there is some difficulty in establishing a single stochastic-process limit that generates useful approximations for both classes. It is natural to let

$$\rho_{i,n} \rightarrow \rho_i , \quad (10.5)$$

where $0 < \rho_i < \infty$. If we let $\rho \equiv \rho_1 + \rho_2 = 1$, then the full system is in heavy traffic, but the high-priority class is in light traffic: $\rho_{1,n} \rightarrow \rho_1 < 1$ as $n \rightarrow \infty$. That implies that the high-priority workload will be asymptotically negligible compared to the total workload in the heavy-traffic scaling. That observation is an important insight, but it does not produce useful approximations for the high-priority class.

On the other hand, if we let $\rho_1 = 1$, then the high-priority class is in heavy traffic, but $\rho \equiv \rho_1 + \rho_2 > 1$, so that the full system is unstable. Clearly, neither of these approaches is fully satisfactory. Yet another approach is to have *both* $\rho_n \rightarrow 1$ and $\rho_{1,n} \rightarrow 1$ as $n \rightarrow \infty$, but that forces $\rho_{2,n} \rightarrow 0$. Such a limit can be useful, but if the low-priority class does not contribute a small proportion of the load, then that approach will usually be unsatisfactory as well.

5.10.1. A Heirarchical Approach

What we suggest instead is a *heirarchical approach* based on considering the relevant scaling. From the scaling analysis in Section 5.5, including the time and space scaling in (5.10) and (5.11), we can see that the full system with higher traffic intensity has greater scaling than the high-priority class alone. Thus, we suggest *first* doing a heavy-traffic stochastic-process limit for the high-priority class alone, based on letting $\rho_{1,n} \uparrow 1$ and, *second*,

afterwards doing a second heavy-traffic limit for both priority classes, based on fixing ρ_1 and letting $\rho_{2,n} \uparrow 1 - \rho_1$.

As a basis for these heavy-traffic limits, we assume that

$$(\mathbf{C}_{1,n}, \mathbf{C}_{2,n}, \mathbf{S}_n) \Rightarrow (\mathbf{C}_1, \mathbf{C}_2, \mathbf{S}) \quad (10.6)$$

where

$$\begin{aligned} \mathbf{C}_{1,n}(t) &\equiv n^{-H_{C,1}}(C_{1,n}(nt) - \lambda_{1,n}nt), \\ \mathbf{C}_{2,n}(t) &\equiv n^{-H_{C,2}}(C_{2,n}(nt) - \lambda_{2,n}nt), \\ \mathbf{S}_n(t) &\equiv n^{-H_S}(S_n(nt) - \mu nt) \end{aligned} \quad (10.7)$$

for $0 < H_{C,1} < 1$, $0 < H_{C,2} < 1$ and $0 < H_S < 1$. For simplicity, we let the processing rate μ be independent of n .

Note that a common case of considerable interest is the light-tailed weak-dependent case with space-scaling exponents

$$H_{C,1} = H_{C,2} = H_S = 1/2, \quad (10.8)$$

but we allow other possibilities. We remark that in the light-tailed case with scaling exponents in (10.8) the heirarchical approach can be achieved directly using strong approximations; see Chen and Shen (2000). (See Section 2.2 of the Internet Supplement for a discussion of strong approximations.)

When (10.8) does not hold, then it is common for one of the three space-scaling exponents to dominate. That leads simplifications in the analysis that should be exploited. In the heavy-traffic limit, variability appears only for the processes with the largest scaling exponent.

Given a heavy-traffic stochastic-process limit as in Theorem 5.4.1 for the high-priority class alone with the space scaling factors in (10.7), we obtain the high-priority approximation

$$W_{1,\rho_1}(t) \approx \left(\frac{\zeta_1}{1 - \rho_1} \right)^{\frac{H_1}{1-H_1}} \mathbf{W}_1 \left(\left(\frac{1 - \rho_1}{\zeta_1} \right)^{\frac{1}{1-H_1}} t \right), \quad t \geq 0, \quad (10.9)$$

as in (5.3) with the scaling functions in (5.10) and (5.11) based on the traffic intensity ρ_1 and the space-scaling exponent

$$H_1 = \max\{H_{C,1}, H_S\}. \quad (10.10)$$

The limit process $\mathbf{W}_1$ in (10.9) is $\phi_K(\mathbf{X}_1)$ as in (4.9), where

$$\mathbf{X}_1(t) = W'_1(0) + \mathbf{C}_1(t) - \mathbf{S}(t) + \eta_1 t, \quad t \geq 0,$$

as in (4.8). If $H_{C,1} > H_S$, then $\mathbf{S}(t) = 0$ in the limit; if $H_S > H_{C,1}$, then $\mathbf{C}_1(t) = 0$ in the limit. Instead of (4.6), here we have

$$\eta_{1,n} \equiv n(\lambda_{1,n} - \mu_n)/c_n \rightarrow \eta_1 .$$

Next we can treat the aggregate workload of both classes using traffic intensity $\rho = \rho_1 + \rho_2$. We can think of the high-priority traffic intensity ρ_1 as fixed with $\rho_1 < 1$ and let $\rho_{2,n} \uparrow 1 - \rho_1$. By the same argument leading to (10.9), we obtain a heavy-traffic stochastic-process limit supporting the approximation

$$W_\rho(t) \approx \left(\frac{\zeta}{1-\rho} \right)^{\frac{H}{1-H}} \mathbf{W} \left(\left(\frac{1-\rho}{\zeta} \right)^{\frac{1}{1-H}} t \right), \quad t \geq 0, \quad (10.11)$$

where the space-scaling exponent now is

$$H = \max\{H_{C,1}, H_{C,2}, H_S\} . \quad (10.12)$$

The limit process $\mathbf{W}$ in (10.11) is $\phi_K(\mathbf{X})$ as in (4.9), where

$$\mathbf{X}(t) = W'(0) + \mathbf{C}_1(t) + \mathbf{C}_2(t) - \mathbf{S}(t) + \eta t, \quad t \geq 0 ,$$

as in (4.8). If $H_{C,i} < H$, then $\mathbf{C}_i(t) = 0$ in the limit; if $H_S < H$, then $\mathbf{S}(t) = 0$ in the limit. Instead of (4.6), here we have

$$\eta_n \equiv n(\lambda_{1,n} + \lambda_{2,n} - \mu_n)/c_n \rightarrow \eta .$$

Not only may the space-scaling exponent H in (10.11) differ from its counterpart H_1 in (10.9), but the parameters ρ and ζ in (10.11) routinely differ from their counterparts ρ_1 and ζ_1 in (10.9).

Of course, the low-priority workload is just the difference between the aggregate workload and the high-priority workload. If that difference is too complicated to work with, we can approximate the low-priority workload by the aggregate workload, since the high-priority workload should be relatively small, i.e.,

$$W_{2,\rho_2}(t) = W_\rho(t) - W_{1,\rho_1}(t) \approx W_\rho(t), \quad t \geq 0 . \quad (10.13)$$

5.10.2. Processing Times

We now consider the *per-class processing times*, i.e., the times required to complete processing of all work of that class in the system. For the

high-priority class, we can apply Theorems 5.9.2 and 5.9.3 to justify (only partially when $K < \infty$) the approximation

$$T_{1,\rho_1}(t) \approx W_{1,\rho_1}(t)/\mu . \quad (10.14)$$

However, the low-priority processing time is more complicated because the last particle of low-priority work must wait, not only for the total aggregate workload to be processed, but also for the processing of all new high-priority work to arrive while that processing of the initial workload is going on. Nevertheless, the low-priority processing time is relatively tractable because it is the time required for the class-1 net input, starting from time t , to decrease far enough to remove the initial aggregate workload, i.e.,

$$T_2(t) \equiv \inf\{u > 0 : X_1(t+u) - X_1(t) < -W(t)\} . \quad (10.15)$$

Note that (10.15) is essentially of the same form as (2.12). Thus, we can apply (10.15) with the reasoning in Theorem 5.9.3 to establish an analog of Theorem 5.9.3, which partly justifies the heavy-traffic approximation

$$T_{2,\rho_1,\rho_2}(t) \approx \frac{W_\rho(t)}{\mu(1-\rho_1)} . \quad (10.16)$$

In (10.16), $T_{2,\rho_1,\rho_2}(t)$ is the low-priority processing time as a function of the two traffic intensities and $W_\rho(t)$ is the aggregate workload at time t as a function of the total traffic intensity $\rho = \rho_1 + \rho_2$.

The heavy-traffic approximation in (10.16) should not be surprising because, as $\rho \uparrow 1$ with ρ_1 fixed, the stochastic fluctuations in X_1 should be negligible in the relatively short time required for the drift in X_1 to hit the target level; i.e., we have a separation of time scales just as in Section 2.4.2.

However, in applications, it may be important to account for the stochastic fluctuations in X_1 . That is likely to be the case when ρ_1 is relatively high compared to ρ . Fortunately, the heavy-traffic limits also suggest a refined approximation. Appropriate heavy-traffic limits for X_1 alone suggest that the stochastic process $\{X_1(t) : t \geq 0\}$ can often be approximated by a Lévy process (a process with stationary and independent increments) without negative jumps. Moreover, the future net input $\{X_1(t+u) - X_1(t) : t \geq 0\}$ often can be regarded as approximately independent of $W(t)$. Under those approximating assumptions, the class-2 processing time in (10.15) becomes tractable. The Laplace transform of the conditional processing-time distribution given $W(t)$ is given on p.120 of Prabhu (1998). The conditional mean is the conditional mean in the heavy-traffic approximation in (10.16).

Remark 5.10.1. *Other service disciplines.* We conclude this section by referring to work establishing heavy-traffic limits for non-FCFS service disciplines. First, in addition to Chen and Shen (2000), Boxma, Cohen and Deng (1999) establish heavy-traffic limits for priority queues. As mentioned in Section 2.4.2, Coffman, Puhalskii and Reiman (1995, 1998), van der Mei and Levy (1997) and van der Mei (2000) establish heavy-traffic limits for polling service disciplines. Kingman (1982) showed how heavy-traffic limits can expose the behavior of a whole class of service disciplines related to random order of service. Yashkov (1993), Sengupta (1992), Grishchkin (1994), Zwart and Boxma (2000) and Boxma and Cohen (2000) establish heavy-traffic limits for the processor-sharing discipline. Fendick and Rodrigues (1991) develop a heavy-traffic approximation for the head-of-the-line generalized processor-sharing discipline. Abate and Whitt (1997a) and Limic (1999) consider the last-in first-out service discipline. Doytchinov et al. (2001) and Kruk et al. (2000) consider “real-time” queues with due dates. These alternative service disciplines are important because they significantly affect queueing performance. As we saw for the high-priority class with two priority classes, the alternative service disciplines can effectively control congestion for some customers when the input of other customers is excessive. The derivations of the heavy-traffic limits with these alternative service disciplines are fascinating because they involve quite different arguments.

Chapter 6

Unmatched Jumps in the Limit Process

6.1. Introduction

As illustrated by the random walks with Pareto steps in Section 1.4 and the workload process with Pareto inputs in Section 2.3, it can be important to consider stochastic-process limits in which the limit process has jumps, i.e., has discontinuous sample paths. The jumps observed in the plots in Chapter 1 correspond to exceptionally large increments in the plotted sequences, i.e., large steps in the simulated random walk and large inputs of required work in the simulated workload process of the queue. Thus, in the associated stochastic-process limit, the jumps in the limit process are *matched* by corresponding jumps in the converging processes. However, there are related situations in which the jumps in the limit process are not matched by jumps in the converging processes.

Indeed, a special focus of this book is on stochastic-process limits with unmatched jumps in the limit process. In the extreme case, the converging stochastic processes have continuous sample paths. Then the sample paths of the converging processes have portions with steep slope corresponding to the limiting jumps. In other cases, a single jump in the sample path of the limiting stochastic process corresponds to many small jumps in the sample path of one of the converging stochastic processes. In this chapter we give several examples showing how a stochastic-process limit with unmatched jumps in the limit process can arise. Most of these examples will be treated in detail later.

We give special attention to stochastic-process limits with unmatched

jumps in the limit process because they represent an interesting phenomenon and because they require special treatment beyond the conventional theory. In particular, as discussed in Section 3.3, whenever there are unmatched jumps in the limit process, we cannot have a stochastic-process limit in the function space D with the conventional Skorohod (1956) J_1 topology. To establish the stochastic-process limit, we instead use the M topology.

Just as in Chapter 1, we primarily draw our conclusions in this chapter by looking at pictures. By plotting initial segments of the stochastic processes for various sample sizes, we can see the stochastic-process limits emerging before our eyes. As before, the plots often do the proper scaling automatically, and thus reveal statistical regularity associated with a macroscopic view of uncertainty. The plots also show the relevance of stochastic-process limits with unmatched jumps in the limit process.

First, though, we should recognize that it is common for the limit process in a stochastic-process limit to have continuous sample paths. For example, that is true for Brownian motion, which is the canonical limiting stochastic process, occurring as the limit in Donsker's theorem, discussed in Chapters 1 and 4. In many books on stochastic-process limits, *all* the stochastic-process limits that are considered have limit processes with continuous sample paths, and there is much to consider.

Moreover, when a limit process in a stochastic-process limit does have discontinuous sample paths, the jumps in the limit process are often matched in the converging processes. We have already pointed out that only matched jumps appear in the examples in Chapter 1. Indeed, there is a substantial literature on stochastic-process limits where the limit process may have jumps and those jumps are matched in the converging processes. The extreme-value limits in Resnick (1987) and the many stochastic-process limits in Jacod and Shiryaev (1987) are all of this form.

However, even for the examples in Chapter 1 with limit processes having discontinuous sample paths, we would have stochastic-process limits with unmatched jumps in the limit process if we formed the continuous-time representation of the discrete-time process using linear interpolation, as in (2.1) in Chapter 1. We contend that the linearly interpolated processes should usually be regarded as asymptotically equivalent to the step-function versions used in Chapter 1; i.e., one sequence of scaled processes should converge if and only if the other does, and they should have the same limit process. That asymptotic equivalence is suggested by Figure 1.13, which plots the two continuous-time representations of a random walk with uniform random steps. As the sample size n increases, both versions approach Brownian motion. Indeed, as n increases, the two alternative continuous-

time representations become indistinguishable.

In Section 6.2 we look at more examples of random walks, comparing the linearly interpolated continuous-time representations (which always have continuous sample paths) to the standard step-function representation for the same random-walk sample paths. Now we make this comparison for random walks approaching a limit process with discontinuous sample paths. Just as in Chapter 1, we obtain jumps in the limit process by considering random walks with steps having a heavy-tailed distribution, in particular, a Pareto distribution. As before, the plots reveal statistical regularity. The plots also show that it is natural to regard the two continuous-time representations of scaled discrete-time processes as asymptotically equivalent.

However, the unmatched jumps in the limit process for the random walks in Section 6.2 can be avoided if we use the step-function representation instead of the linearly interpolated version. Since the step-function version seems more natural anyway, the case for considering unmatched jumps in the limit process is not yet very strong. In the rest of this chapter we give examples in which stochastic-process limits with unmatched jumps in the limit process cannot be avoided.

6.2. Linearly Interpolated Random Walks

All the stochastic-process limits with jumps in the limit process considered in Chapter 1 produce unmatched jumps when we form the continuous-time representation of the original discrete-time process by using linear interpolation. We now want to show, by example, that it is natural to regard the linearly interpolated continuous-time representation as asymptotically equivalent to the standard step-function representation in settings where the limit process has jumps.

Given a random walk or any discrete-time process $\{S_k : k \geq 0\}$, the scaled-and-centered step-function representations are defined for each $n \geq 1$ by

$$\mathbf{S}_n(t) \equiv c_n^{-1}(S_{[nt]} - m[nt]), \quad 0 \leq t \leq 1, \quad (2.1)$$

where $[x]$ is the greatest integer less than x and $c_n \rightarrow \infty$ as $n \rightarrow \infty$. The associated linearly interpolated versions are

$$\tilde{\mathbf{S}}_n(t) \equiv (nt - [nt])\mathbf{S}_n([nt] + 1/n) + (1 + [nt] - nt)\mathbf{S}_n([nt]/n), \quad (2.2)$$

for $0 \leq t \leq 1$. Clearly the sample paths of $\mathbf{S}_n$ in (2.1) are discontinuous for all n (except in the special case in which $S_k = S_0, 1 \leq k \leq n$), while the sample paths of $\tilde{\mathbf{S}}_n$ in (2.2) are continuous for all n .

6.2.1. Asymptotic Equivalence with M_1

We contend that the two sequences of processes $\{\mathbf{S}_n : n \geq 0\}$ and $\{\tilde{\mathbf{S}}_n : n \geq 0\}$ in the function space $D \equiv D([0, 1], \mathbb{R})$ should be *asymptotically equivalent*, i.e., if either converges in distribution as $n \rightarrow \infty$, then so should the other, and they should have the same limit. It is easy to see that the desired asymptotic equivalence holds with the M_1 metric. In particular, we can show that $d_{M_1}(\mathbf{S}_n, \tilde{\mathbf{S}}_n) \Rightarrow 0$ as $n \rightarrow \infty$.

Theorem 6.2.1. (the M_1 distance between the continuous-time representations) *For any discrete-time process $\{S_k : k \geq 0\}$,*

$$d_{M_1}(\mathbf{S}_n, \tilde{\mathbf{S}}_n) \leq n^{-1} \quad \text{for all } n \geq 1,$$

for $\mathbf{S}_n$ in (2.1) and $\tilde{\mathbf{S}}_n$ in (2.2).

Proof. For the M_1 metric, we can use an arbitrary parametric representation of the step-function representation $\mathbf{S}_n$. Then, for any $\epsilon > 0$, we can construct the associated parametric representation of $\tilde{\mathbf{S}}_n$ so that it agrees with the other parametric representation at the finitely many points in the domain $[0, 1]$ mapping into the points $(k/n, \mathbf{S}_n(k/n))$ on the completed graph of $\mathbf{S}_n$ for $0 \leq k \leq n$, with the additional property that the spatial components of the two parametric representations differ by at most $n^{-1} + \epsilon$ anywhere. Since ϵ was arbitrary, we obtain the desired conclusion. ■

We can apply Theorem 6.2.1 and the convergence-together theorem, Theorem 11.4.7, to establish the desired asymptotic equivalence with respect to convergence in distribution.

Corollary 6.2.1. (asymptotic equivalence of continuous-time representations) *If either $\mathbf{S}_n \Rightarrow \mathbf{S}$ in (D, M_1) or $\tilde{\mathbf{S}}_n \Rightarrow \mathbf{S}$ in (D, M_1) , then both limits hold.*

Note that the conclusion of Theorem 6.2.1 is much stronger than the conclusion of Corollary 6.2.1. Corollary 6.2.1 concludes that $\mathbf{S}_n$, $\tilde{\mathbf{S}}_n$ and $\mathbf{S}$ all have approximately the same probability laws for all suitably large n , whereas Theorem 6.2.1 concludes that the individual sample paths of $\mathbf{S}_n$ and $\tilde{\mathbf{S}}_n$ are likely to be close for all suitably large n .

We used plots to illustrate the asymptotic equivalence of $\tilde{\mathbf{S}}_n$ and $\mathbf{S}_n$ for random walks with uniform steps, for which the limit process is Brownian motion, in Figure 1.13. That asymptotic equivalence is proved by Corollary 6.2.1. (Since the limit process has continuous sample paths, the various non-uniform Skorohod topologies are equivalent in this example.)

Now we use plots again to illustrate the asymptotic equivalence of $\tilde{\mathbf{S}}_n$ and $\mathbf{S}_n$ in random walks with jumps in the limit process. Since the asymptotic equivalence necessarily holds in the M_1 topology by virtue of Corollary 6.2.1, but not in the J_1 topology, we are presenting a case for using the M_1 topology.

6.2.2. Simulation Examples

We give three examples, all involving variants of the Pareto distribution.

Example 6.2.1. *Centered random walk with Pareto(p) steps.*

As in (3.5) (iii) in Section 1.3, we consider the random walk $\{S_k : k \geq 0\}$ with IID steps

$$X_k \equiv U_k^{-1/p} \quad (2.3)$$

for U_k uniformly distributed on the interval $[0, 1]$. The steps then have a Pareto(p) distribution with parameter p , having ccdf $F^c(t) = t^{-p}$ for $t \geq 1$. We first consider the case $1 < p \leq 2$. In that case, the steps have a finite mean $m = 1 + (p - 1)^{-1}$ but infinite variance. In Figures 1.20 – 1.22, we saw that the plots of the centered random walks give evidence of jumps. The supporting FCLT (in Section 4.5) states that the step-function representations converge in distribution to a stable Lévy motion, which indeed has discontinuous sample paths.

Just as in Chapter 1, we use the statistical package S to simulate and plot the initial segments of the stochastic processes. Plots of the two continuous-time representations $\mathbf{S}_n$ and $\tilde{\mathbf{S}}_n$ for the same sample paths of the random walk are given for the case $p = 1.5$ and $n = 10^j$ with $j = 1, 2, 3$ in Figure 6.1. For $n = 10$, the two continuous-time representations look quite different. Indeed, at first it may seem that they cannot be corresponding continuous-time representations of the same realized segment of the random walk, but closer examination shows that the two continuous-time representations are correct. However, for $n = 100$ and beyond, the two continuous-time representations look very similar. For larger values of n such as $n = 10^4$ and beyond, the two continuous-time representations look virtually identical.

So far we have considered only $p = 1.5$. We now illustrate how the plots depend on p for $1 < p \leq 2$. In Figure 6.2 we plot the two continuous-time representations of the random walk with Pareto(p) steps for three values of p , in particular for $p = 1.1, 1.5$ and 1.9 . We do the plot for the case $n = 100$ using the *same uniform random numbers* (exploiting (2.3)). In each plot the largest steps stem from the smallest uniform random numbers. The

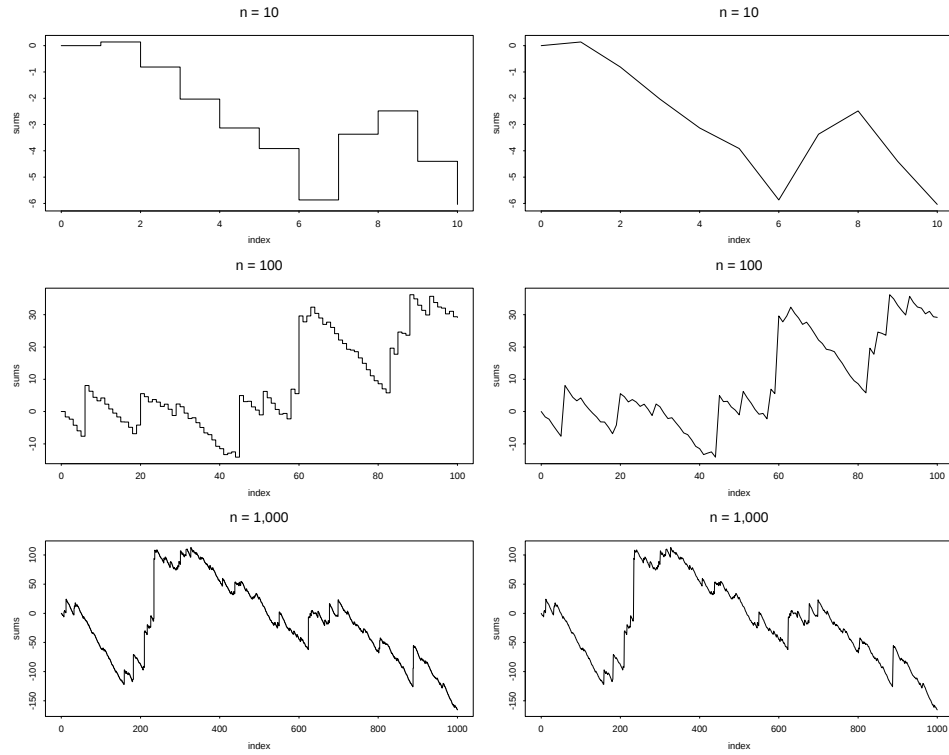


Figure 6.1: Plots of the two continuous-time representations of the centered random walk with Pareto(1.5) steps for $n = 10^j$ with $j = 1, 2, 3$. The step-function representation $\mathbf{S}_n$ in (2.1) appears on the left, while the linearly interpolated version $\tilde{\mathbf{S}}_n$ in (2.2) appears on the right.

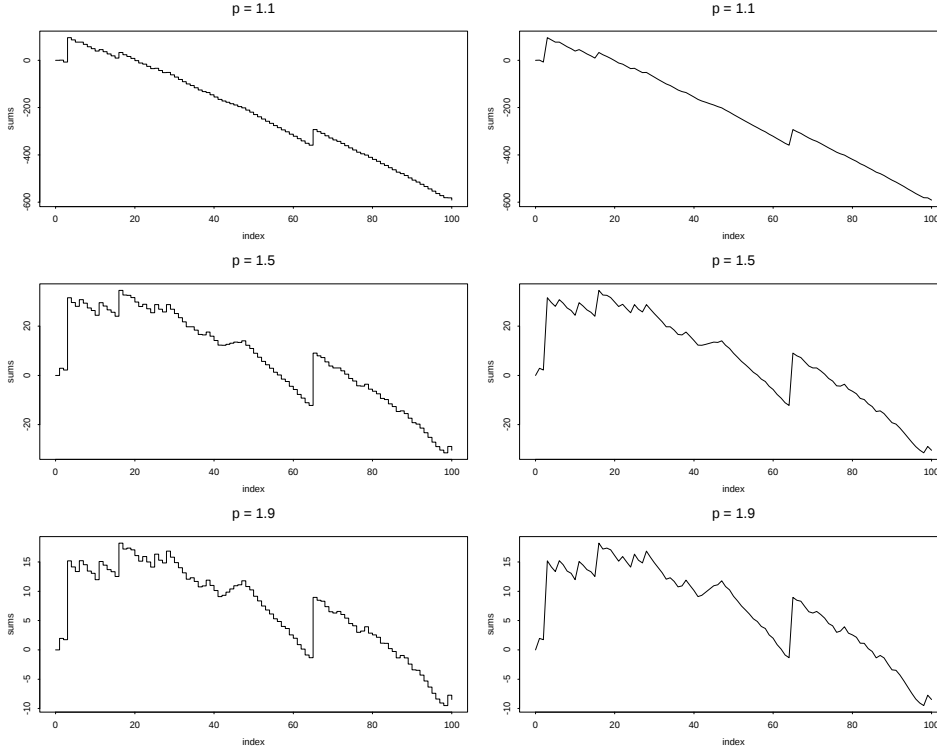
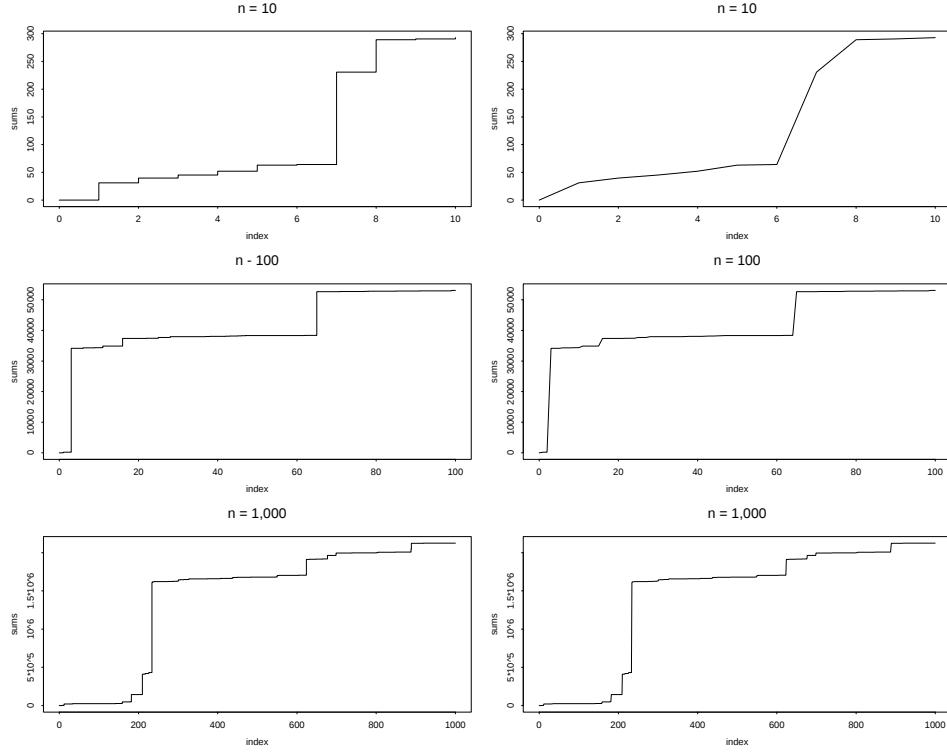


Figure 6.2: Plots of the two continuous-time representations of the centered random walk with $\text{Pareto}(p)$ steps with $p = 1.1, 1.5$ and 1.9 for $n = 10^2$ based on the same uniform random numbers (using (2.3)). The step-function representation $\mathbf{S}_n$ in (2.1) appears on the left, while the linearly interpolated version $\tilde{\mathbf{S}}_n$ in (2.2) appears on the right.

three smallest uniform random numbers in this sample were $U_3 = 0.00542$, $U_{65} = 0.00836$ and $U_{16} = 0.0201$. The corresponding large steps can be seen in each case of Figure 6.2. Again, we see that the limiting stochastic process should have jumps (up). That conclusion is confirmed by considering larger and larger values of n . As in Figures 6.1 and 6.2, the two continuous-time representations look very similar. And the little difference we see for $n = 100$ decreases as n increases.

Example 6.2.2. *Uncentered random walk with $\text{Pareto}(0.5)$ steps.* In Figures 1.19, 1.25 and 1.26 we saw that the *uncentered* random walk with $\text{Pareto}(0.5)$ steps should have stochastic-process limits with jumps in the limit process. The supporting FCLT implies convergence to another stable



Plots of the uncentered random walk with Pareto(0.5) steps for $n = 10^j$ with $j = 1, 2, 3$. The step-function representation $\mathbf{S}_n$ in (2.1) appears on the left, while the linearly interpolated version $\tilde{\mathbf{S}}_n$ in (2.2) appears on the right.

Lévy motion as $n \rightarrow \infty$ (again see Section 4.5). Moreover, such a limit holds for IID Pareto(p) steps whenever $p \leq 1$, because then the steps have infinite mean.

Now we look at the two continuous-time representations in this setting. We now plot the two continuous-time representations $\tilde{\mathbf{S}}_n$ and $\mathbf{S}_n$ associated with the uncentered random walk with Pareto(0.5) steps for $n = 10^j$ with $j = 1, 2, 3$ in Figure 6.2.2. Again, the two continuous-time representations initially (for small n) look quite different, but become indistinguishable as n increases. Just as in Chapter 1, even though there are jumps, we see statistical regularity associated with large n . Experiments with different n show the self-similarity discussed before.

Example 6.2.3. *Centered random walk with limiting jumps up and down.*

The Pareto distributions considered above have support on the inter-

val $[1, \infty)$, so that, even with centering, the positive tail of the step-size distribution is heavy, but the negative tail of the step-size distribution is light. Consequently the limiting stochastic process in the stochastic-process limit for the random walks with Pareto steps can only have jumps up. (See Section 4.5)

We can obtain a limit process with both jumps up and jumps down if we again use (2.3) to define the steps, but we let U_k be uniformly distributed on the interval $[-1, 1]$ instead of in $[0, 1]$. Then we can have both arbitrarily large negative jumps and arbitrarily large positive jumps. We call the resulting distribution a *symmetric Pareto distribution* (with parameter p). Since the distribution is symmetric, no centering need be done for the plots or the stochastic-process limits.

To illustrate, we make additional comparisons between the linearly interpolated continuous-time representation and the step-function continuous-time representation of the random walk, now using the symmetric Pareto(p) steps for $p = 1.5$. The plots are shown in Figure 6.3. We plot the two continuous-time representations for $n = 10^j$ with $j = 2, 3, 4$. From the plots, it is evident that the limit process now should have jumps down as well as jumps up. Again, the two continuous-time representations look almost identical for large n .

6.3. Heavy-Tailed Renewal Processes

One common setting for stochastic-process limits with unmatched jumps in the limit process, which underlies many applications, is a heavy-tailed renewal process. Given partial sums $S_k \equiv X_1 + \cdots + X_k, k \geq 1$, from a sequence of nonnegative random variables $\{X_k : k \geq 1\}$ (without an IID assumption), the associated stochastic process $N \equiv \{N(t) : t \geq 0\}$ defined by

$$N(t) \equiv \max\{k \geq 0 : S_k \leq t\}, \quad t \geq 0, \quad (3.1)$$

where $S_0 \equiv 0$, is called a *stochastic counting process*. When the random variables X_k are IID, the counting process is called a *renewal counting process* or just a *renewal process*.

6.3.1. Inverse Processes

Roughly speaking (we will be more precise in Chapter 13), the stochastic processes $\{S_k : k \geq 1\}$ and $N \equiv \{N(t) : t \geq 0\}$ can be regarded as *inverses*

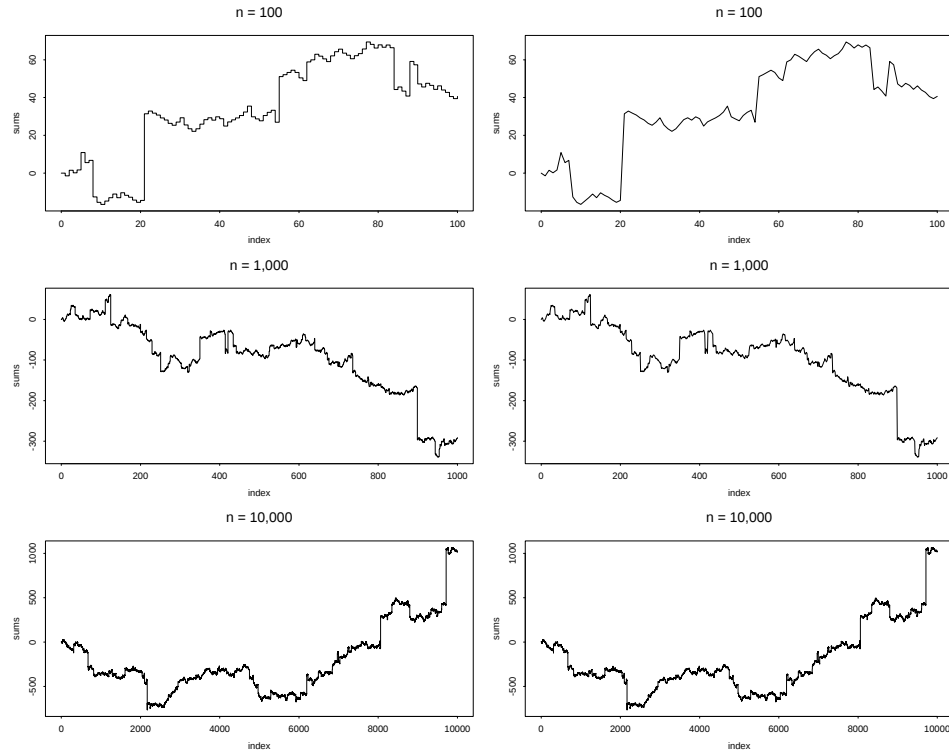


Figure 6.3: Plots of the two continuous-time representations of the random walk with symmetric Pareto(1.5) steps for $n = 10^j$ with $j = 2, 3, 4$. The step-function representation $\mathbf{S}_n$ in (2.1) appears on the left, while the linearly interpolated version $\tilde{\mathbf{S}}_n$ in (2.2) appears on the right.

of each other, without imposing the IID condition, because

$$S_k \leq t \quad \text{if and only if} \quad N(t) \geq k. \quad (3.2)$$

The M_1 topology is convenient for relating limits for partial sums to associated limits for the counting processes, because the M_1 -topology definition makes it easy to exploit the inverse relation in the continuous-mapping approach.

Moreover, it is not possible to use the standard J_1 topology to establish limits of scaled versions of the counting processes, because the J_1 topology requires all jumps in the limit process to be matched in the converging stochastic processes. The difficulty with the J_1 topology on D can easily be seen when the random variables X_k are strictly positive. Then the counting process N increases in unit jumps, and scaled versions of the counting process, such as

$$\mathbf{N}_n(t) \equiv c_n^{-1}(N(nt) - m^{-1}nt), \quad t \geq 0, \quad (3.3)$$

where $c_n \rightarrow \infty$, have jumps of magnitude $1/c_n$, which are asymptotically negligible as $n \rightarrow \infty$. Hence, if $\mathbf{N}_n$ in (3.3) is ever to converge as $n \rightarrow \infty$ to a limiting stochastic process with discontinuous sample paths, then we must have unmatched jumps in the limit process. Then we need the M_1 topology on D .

What is not so obvious, however, is that $\mathbf{N}_n$ will ever converge to a limiting stochastic process with discontinuous sample paths. However, such limits can indeed occur. Here is how: A long interrenewal time creates a long interval between jumps up in the renewal process. The long interrenewal time appears horizontally rather than vertically, not directly causing a jump. However, during such an interval, the scaled process in (3.3) will decrease linearly at rate n/mc_n , due to the translation term not being compensated for by any jumps up. When $n/c_n \rightarrow \infty$ (the usual case), the slope approaches $-\infty$. When the interrenewal times are long enough, these portions of the sample path with steep slope down can lead to jumps *down* in the limit process.

A good way to see how jumps can appear in the limit process for $\mathbf{N}_n$ is to see how limits for $\mathbf{N}_n$ in (3.3) are related to associated limits for $\mathbf{S}_n$ in (2.1) when both scaled processes are constructed from the same underlying process $\{S_k : k \geq 0\}$. A striking result from the continuous-mapping approach to stochastic-process limits (to be developed in Chapter 13) is an equivalence between stochastic-process limits for partial sums and associated counting processes, exploiting the M_1 topology (but not requiring any

direct independence or common-distribution assumption). As a consequence of Corollary 13.8.1, we have the following result:

Theorem 6.3.1. (FCLT equivalence for counting processes and associated partial sums) *Suppose that $0 < m < \infty$, $c_n \rightarrow \infty$, $n/c_n \rightarrow \infty$ and $\mathbf{S}(0) = 0$. Then*

$$\mathbf{S}_n \Rightarrow \mathbf{S} \quad \text{in} \quad (D, M_1) \quad (3.4)$$

for $\mathbf{S}_n$ in (2.1) if and only if

$$\mathbf{N}_n \Rightarrow \mathbf{N} \quad \text{in} \quad (D, M_1) \quad (3.5)$$

for $\mathbf{N}_n$ in (3.3), in which case

$$(\mathbf{S}_n, \mathbf{N}_n) \Rightarrow (\mathbf{S}, \mathbf{N}) \quad \text{in} \quad (D^2, WM_1), \quad (3.6)$$

where the limit processes are related by

$$\mathbf{N}(t) \equiv (m^{-1}\mathbf{S} \circ m^{-1}\mathbf{e})(t) \equiv m^{-1}\mathbf{S}(m^{-1}t), \quad t \geq 0, \quad (3.7)$$

or, equivalently,

$$\mathbf{S}(t) = (m\mathbf{N} \circ m\mathbf{e})(t) \equiv m\mathbf{N}(mt), \quad t \geq 0, \quad (3.8)$$

where $\mathbf{e}(t) = t$, $t \geq 0$.

Thus, whenever the limit process $\mathbf{S}$ in (3.4) has discontinuous sample paths, the limit process $\mathbf{N}$ in (3.5) necessarily has discontinuous sample paths as well. Moreover, $\mathbf{S}$ has only jumps up (down) if and only if $\mathbf{N}$ has only jumps down (up). Whenever $\mathbf{S}$ and $\mathbf{N}$ have discontinuous sample paths, the M_1 topology is needed to express the limit for $\mathbf{N}_n$ in (3.5). In contrast, the limit for $\mathbf{S}_n$ in (3.4) can hold in (D, J_1) .

6.3.2. The Special Case with $m = 1$

The close relation between the limit processes $\mathbf{S}$ and $\mathbf{N}$ in (3.4) – (3.8) is easy to understand and visualize when we consider plots for the special case of strictly positive steps X_k with translation scaling constant $m = 1$. Note that the limit process $\mathbf{N}$ in (3.7) becomes simply $-\mathbf{S}$ when $m = 1$.

Also note that we can always scale so that $m = 1$ without loss of generality: For any given sequence $\{X_k : k \geq 0\}$, when we multiply X_k by m for all k , we replace $\mathbf{S}_n$ by $m\mathbf{S}_n$ and $\mathbf{N}_n$ by $\mathbf{N}_n \circ m^{-1}\mathbf{e}$. Hence, the limits $\mathbf{S}$ and $\mathbf{N}$ are replaced by $m\mathbf{S}$ and $\mathbf{N} \circ m^{-1}\mathbf{e}$, respectively.

Hence, suppose that $m = 1$. A useful observation, then, is that $N(S_k) = k$ for all k . (We use the assumption that the variables X_k are strictly positive.) With that in mind, note that we can plot $N(t) - t$ versus t , again using the statistical package *S*, by plotting the points $(0, 0)$, $(S_k, N(S_k) - 1 - S_k)$ and $(S_k, N(S_k) - S_k)$ in the plane $\mathbb{R}^2$ and then performing linear interpolation between successive points.

Roughly speaking, then, we can plot $N(t) - t$ versus t by plotting $N(S_k) - S_k$ versus S_k . On the other hand, when we plot the centered random walk $\{S_k - k : k \geq 0\}$, we plot $(S_k - k)$ versus k . Since $N(S_k) = k$, we have

$$N(S_k) - S_k = k - S_k = -(S_k - k) .$$

Thus, the second component of the pair $(S_k, N(S_k) - S_k)$ is just minus 1 times the second component of the pair $(k, S_k - k)$. Thus, the plot of $N(t) - t$ versus t should be very close to the plot of $-(S_k - k)$ versus k . The major difference is in the first component: For the renewal process, the first component is S_k ; for the random walk, the first component is k . However, since $n^{-1}S_n \rightarrow 1$ as $n \rightarrow \infty$ by the SLLN, that difference between these two first components disappears as $n \rightarrow \infty$.

Example 6.3.1. *Centered renewal processes with Pareto(p) steps for $1 < p < 2$.* By now, we are well acquainted with a situation in which the limit for $\mathbf{S}_n$ in (3.4) holds and the limit process $\mathbf{S}$ has discontinuous sample paths: That occurs when the underlying process $\{S_k : k \geq 0\}$ is a random walk with IID Pareto(p) steps for $1 < p < 2$. Then the limit (3.4) holds with $m = 1 + (p - 1)^{-1}$ and $\mathbf{S}$ being a stable Lévy motion, which has discontinuous sample paths. The discontinuous sample paths are clearly revealed for the case $p = 1.5$ in Figures 1.20 – 1.22 and 6.1.

To make the relationship clear, we consider the case $m = 1$. We obtain $m = 1$ in our example with IID Pareto(1.5) steps by dividing the steps by 3; i.e., we let $X_k \equiv U_k^{-2/3}/3$. For this example with Pareto(1.5) steps having cdf decay rate $p = 3/2$ and mean 1, we plot both the centered renewal process ($N(t) - t$ versus t) and minus 1 times the centered random walk ($-(S_k - k)$ versus k). We plot both sample paths, putting the centered renewal process on the left, for the cases $n = 10^j$ with $j = 1, 2, 3$ in Figure 6.4. We plot three possible representations of each for $n = 10^4$ in Figure 6.5. (We plot the centered random walk directly; i.e., we do not use either of the continuous-time representations.)

For small n , the sample paths of the two centered processes look quite different, but as n increases, the sample paths begin to look alike. The

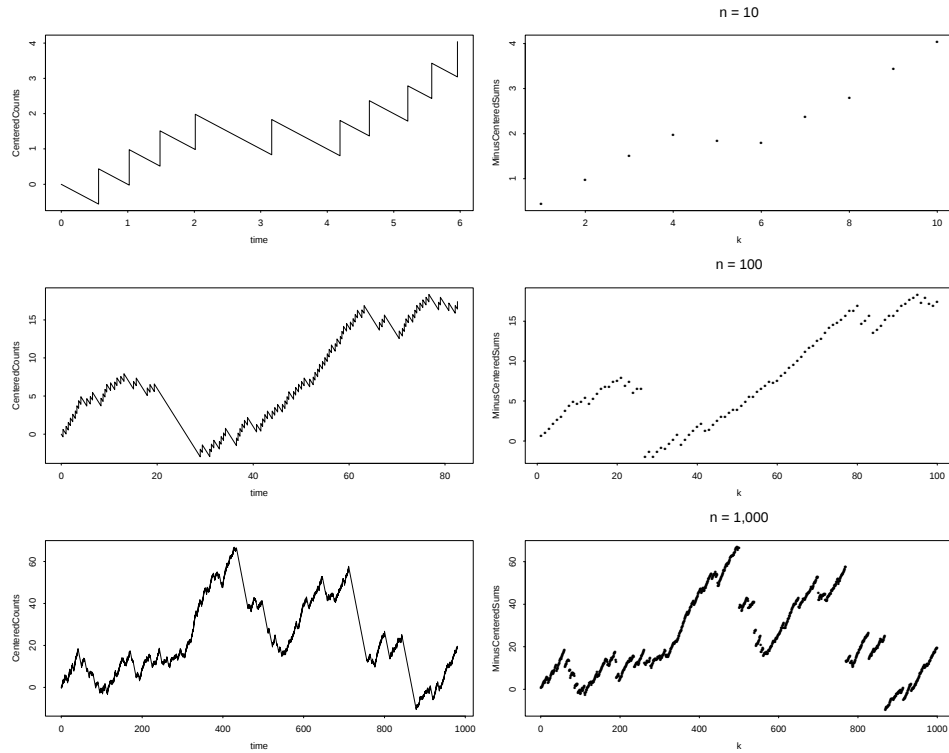


Figure 6.4: Plots of the centered renewal process (on the left) and minus 1 times the centered random walk (on the right) for Pareto(1.5) steps with mean $m = 1$ and $n = 10^j$ for $j = 1, 2, 3$.

jumps in the centered random walk plot are matched with portions of the centered-renewal-process plot with very steep slope. As n increases, the slopes in the portions of the centered-renewal-process plots corresponding to the random-walk jumps tend to get steeper and steeper, approaching the jump itself.

It is natural to wonder how the plots look as the decay rate p changes within the interval $(1, 2)$, which is the set of values yielding a finite mean but an infinite variance. We know that for smaller p the jumps are likely to be larger. To see what happens, we plot three realizations each of the centered renewal process and minus 1 times the centered random walk for Pareto steps having decay rates $p = 7/4$ and $p = 5/4$ (normalized as before to have mean 1) for $n = 10^4$ in Figures 6.6 and 6.7. From Figures 6.5 – 6.7, we see that the required space scaling decreases, the two irregular paths

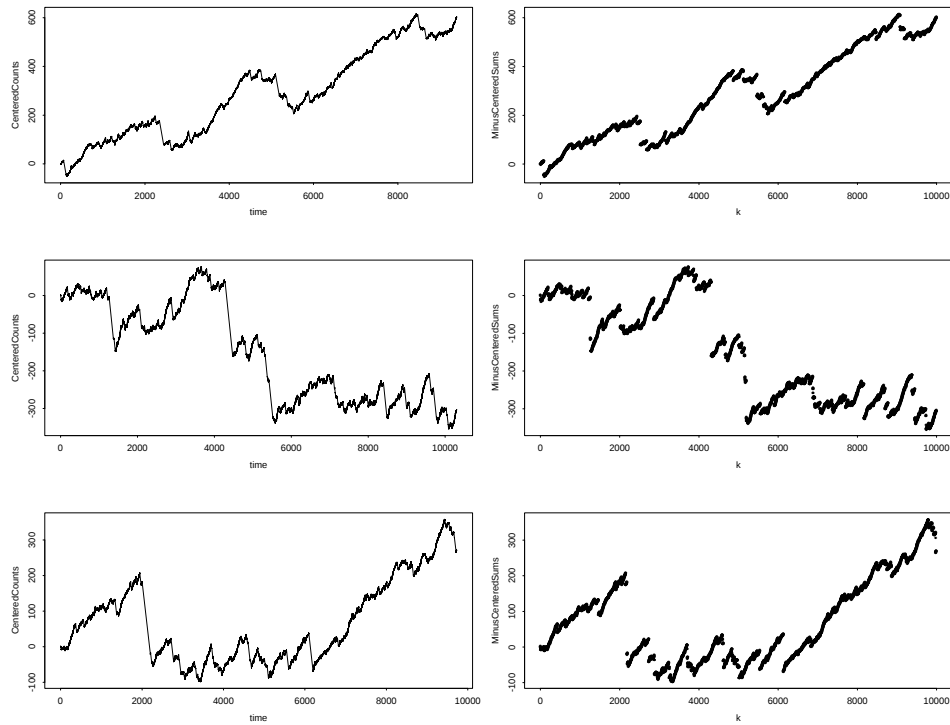


Figure 6.5: Plots of three independent realizations of the centered renewal process (on the left) and minus 1 times the centered random walk (on the right) for Pareto(1.5) steps with mean $m = 1$ and $n = 10^4$.

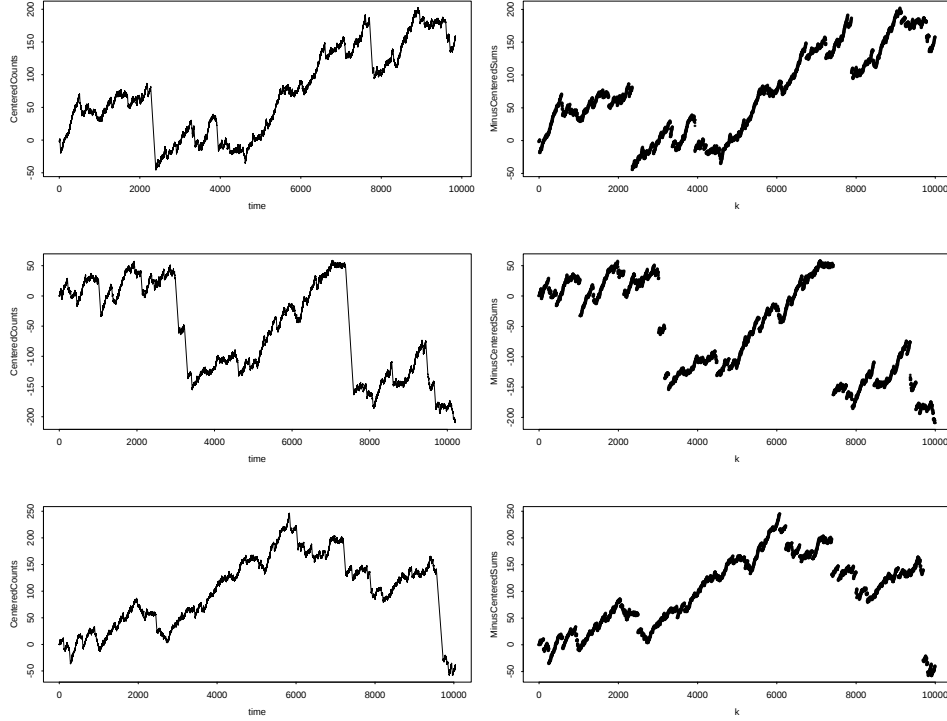


Figure 6.6: Plots of three independent realizations of the centered renewal process (on the left) and minus 1 times the centered random walk (on the right) associated with Pareto(p) steps in (2.3) with $p = 7/4$, $m = 1$ and $n = 10^4$.

become closer, and the slopes in the renewal-process plot become steeper, as p increases from $5/4$ to $3/2$ to $7/4$. For $p = 5/4$, we need larger n to see steeper slopes. However, in all cases we can see that there should be unmatched jumps in the limit process. ■

For the Pareto-step random walk plots in Figures 6.4 – 6.7, we not only have $-\mathbf{S}_n \Rightarrow -\mathbf{S}$ and $\mathbf{N}_n \Rightarrow -\mathbf{S}$, but also the realizations of $\mathbf{N}_n$ and $-\mathbf{S}_n$ are becoming close to each other as $n \rightarrow \infty$. Such asymptotic equivalence follows from Theorem 6.3.1 by virtue of Theorem 11.4.8. Recall that we can start with any translation scaling constant m and rescale to $m = 1$.

Corollary 6.3.1. (asymptotic equivalence) *If, in addition to the assump-*

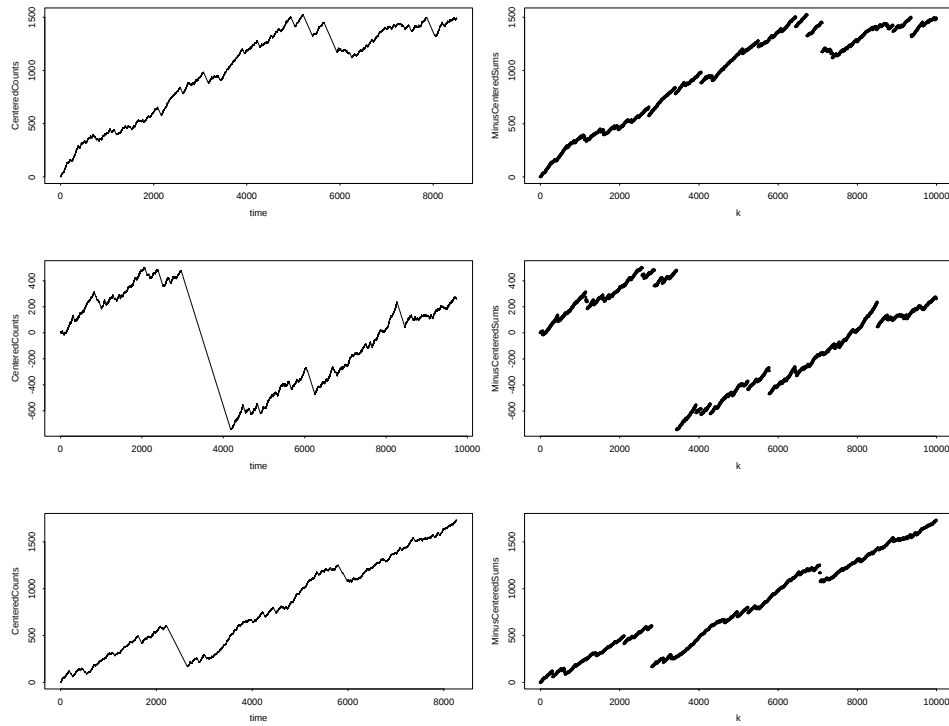


Figure 6.7: Plots of three independent realizations of the centered renewal process (on the left) and minus 1 times the centered random walk (on the right) associated with $\text{Pareto}(p)$ steps in (2.3) with $p = 5/4$, $m = 1$ and $n = 10^4$.

tions of Theorem 6.3.1, the limit $\mathbf{S}_n \Rightarrow \mathbf{S}$ in (3.4) holds and $m = 1$, then

$$d_{M_1}(\mathbf{N}_n, -\mathbf{S}_n) \Rightarrow 0 .$$

To summarize, properly scaled versions (with centering) of a renewal process (or, more generally, any counting process) are intimately connected with associated scaled versions (with centering) of random walks, so that FCLTs for random walks imply associated FCLTs for the scaled renewal process (and vice versa), provided that we use the M_1 topology. When the limit process for the random walk has discontinuous sample paths, so does the limit process for the renewal process, which necessarily produces unmatched jumps. We state specific FCLTs for renewal processes in Section 7.3.

6.4. A Queue with Heavy-Tailed Distributions

Closely paralleling the heavy-tailed renewal process just considered, heavy-traffic limits for the queue-length process in standard queueing models routinely produce stochastic-process limits with unmatched jumps in the limit process when the service times or interarrival times have heavy-tailed distributions (again meaning with infinite variance). In fact, renewal processes enter in directly, because the customer arrival process in the queueing model is a stochastic counting process, which is a renewal process when the interarrival times are IID.

We start by observing that jumps in the limit process associated with stochastic-process limits for the queue-length process almost always are unmatched jumps. That is easy to see when all the interarrival times and service times are strictly positive. (That is the case w.p.1 when the interarrival times and service times come from sequences of random variables with distributions assigning 0 probability to 0.) Then the queue length (i.e., the number of customers in the system) makes changes in unit steps. Thus, any jumps in the limit process associated with a stochastic-process limit for a sequence of queue-length processes with space scaling, where we divide by c_n with $c_n \rightarrow \infty$ as $n \rightarrow \infty$, must be unmatched jumps.

The real issue, then, is to show that jumps can appear in stochastic-process limits for the queue-length process. The stochastic-process limits we have in mind occur in a heavy-traffic setting, as in Section 2.3.

6.4.1. The Standard Single-Server Queue

To be specific, we consider a single-server queue with unlimited waiting room and the first-come first-served service discipline. (We will discuss this model further in Chapter 9. The model can be specified by a sequence of ordered pairs of nonnegative random variables $\{(U_k, V_k) : k \geq 1\}$. The variable U_k represents the *interarrival time* between customers k and $k - 1$, with U_1 being the arrival time of the first customer, while the variable V_k represents the *service time* of the customer k . The *arrival time* of the customer k is thus

$$T_k \equiv U_1 + \cdots + U_k, \quad k \geq 1, \quad (4.1)$$

and the *departure time* of the customer k is

$$D_k \equiv T_k + W_k + V_k, \quad k \geq 1, \quad (4.2)$$

where W_k is the *waiting time* (before beginning service) of customer k . The waiting times can be defined recursively by

$$W_k \equiv [W_{k-1} + V_{k-1} - U_k]^+, \quad k \geq 2, \quad (4.3)$$

where $[x]^+ \equiv \max\{x, 0\}$ and $W_1 \equiv 0$. (We have assumed that the system starts empty; that of course is not critical.)

We can now define associated continuous-time processes. The counting processes are defined just as in (3.1). The *arrival (counting) process* $\{A(t) : t \geq 0\}$ is defined by

$$A(t) \equiv \max\{k \geq 0 : T_k \leq t\}, \quad t \geq 0, \quad (4.4)$$

the *departure (counting) process* $\{D(t) : t \geq 0\}$ is defined by

$$D(t) \equiv \max\{k \geq 0 : D_k \leq t\}, \quad t \geq 0, \quad (4.5)$$

and the *queue-length process* $\{Q(t) : t \geq 0\}$ is defined by

$$Q(t) \equiv A(t) - D(t), \quad t \geq 0. \quad (4.6)$$

Here the queue length is the number in system, including the customer in service, if any.

The standard single-server queue that we consider now is closely related to the infinite-capacity version of the discrete-time fluid queue model considered in Section 2.3. Indeed, the recursive definition for the waiting times in (4.3) is essentially the same as the recursive definition for the workloads

in (3.1) of Section 2.3 in the special case in which the waiting space is unlimited, i.e., when $K = \infty$. For the fluid queue model, we saw that the behavior of the workload process is intimately connected to the behavior of an associated random walk, and that heavy-tailed inputs lead directly to jumps in the limit process for appropriately scaled workload processes. The same is true for the waiting times here, as we will show in Section 9.2.

6.4.2. Heavy-Traffic Limits

Thus, just as in Section 2.3, we consider a sequence of models indexed by n in order to obtain interesting stochastic-process limits for stable queueing systems. We can achieve such a framework conveniently by scaling a single model. We use a superscript n to index the new quantities constructed in the n^{th} model.

We start with a single sequence $\{(U_k, V_k) : k \geq 1\}$. Note that we have made no stochastic assumptions so far. The key assumption is a FCLT for the random walks, in particular,

$$(\mathbf{S}_n^u, \mathbf{S}_n^v) \Rightarrow (\mathbf{S}^u, \mathbf{S}^v) \quad \text{in} \quad (D^2, WM_1), \quad (4.7)$$

where

$$\mathbf{S}_n^u \equiv c_n^{-1} \left(\sum_{i=1}^{\lfloor nt \rfloor} U_i - \lfloor nt \rfloor \right)$$

and

$$\mathbf{S}_n^v \equiv c_n^{-1} \left(\sum_{i=1}^{\lfloor nt \rfloor} V_i - \lfloor nt \rfloor \right).$$

The standard stochastic assumption to obtain (4.7) is for $\{U_k\}$ and $\{V_k\}$ to be independent sequences of IID random variables with

$$EV_k = EU_k = 1 \quad \text{for all} \quad k \geq 1. \quad (4.8)$$

and other regularity conditions (finite variances to get convergence to Brownian motion or asymptotic power tails to get convergence to stable Lévy motions).

Paralleling the scaling in (3.13) in Section 2.3, we form the n^{th} model by letting

$$U_k^n \equiv b_n U_k \quad \text{and} \quad V_k^n \equiv V_k, \quad k \geq 1, \quad (4.9)$$

where

$$b_n \equiv 1 + mc_n/n \quad \text{for} \quad n \geq 1. \quad (4.10)$$

We assume that $c_n/n \downarrow 0$ as $n \rightarrow \infty$, so that $b_n \downarrow 1$ as $n \rightarrow \infty$. The scaling in (4.9) is a simple deterministic scaling of time in the arrival process; i.e., the arrival process in model n is

$$A^n(t) \equiv A(b_n^{-1}t), \quad t \geq 0 ,$$

for b_n in (4.10).

We now form scaled stochastic processes associated with the sequence of models by letting

$$\mathbf{W}_n(t) \equiv c_n^{-1}W_{[nt]}^n, \quad (4.11)$$

and

$$\mathbf{Q}_n(t) \equiv c_n^{-1}Q^n(nt), \quad t \geq 0 . \quad (4.12)$$

We now state the heavy-traffic stochastic-process limit, which follows from Theorems 9.3.3, 9.3.4 and 11.4.8. As before, for $x \in D$, let $\text{Disc}(x)$ be the set of discontinuities of x .

Theorem 6.4.1. (heavy-traffic limit for the waiting times and queue lengths)
Suppose that the stochastic-process limit in (4.7) holds and the scaling in (4.9) holds with $c_n \rightarrow \infty$ and $c_n/n \rightarrow 0$. Suppose that almost surely the sets $\text{Disc}(\mathbf{S}^u)$ and $\text{Disc}(\mathbf{S}^v)$ have empty intersection and

$$P(\mathbf{S}^u(0) = 0) = P(\mathbf{S}^v(0) = 0) = 1 .$$

Then

$$\mathbf{W}_n \Rightarrow \mathbf{W} \equiv \phi(\mathbf{S}^v - \mathbf{S}^u - m\mathbf{e}) \quad \text{in } (D, M_1) , \quad (4.13)$$

where ϕ is the one-sided reflection map in (5.4) in Section 3.5,

$$(\mathbf{W}_n, \mathbf{Q}_n) \Rightarrow (\mathbf{W}, \mathbf{W}) \quad \text{in } (D^2, WM_1) \quad (4.14)$$

and

$$d_{M_1}(\mathbf{W}_n, \mathbf{Q}_n) \Rightarrow 0 . \quad (4.15)$$

We now explain why the limit process $\mathbf{Q}$ for the scaled queue-length processes can have jumps. Starting from (4.6), we have

$$\mathbf{Q}_n = \mathbf{A}_n - \mathbf{D}_n , \quad (4.16)$$

where

$$\mathbf{A}_n(t) \equiv c_n^{-1}(A^n(nt) - nt), \quad t \geq 0 \quad (4.17)$$

and

$$\mathbf{D}_n(t) \equiv c_n^{-1}(D^n(nt) - nt), \quad t \geq 0 . \quad (4.18)$$

Just as for the renewal processes in the previous section, an especially long service time (interarrival time) can cause a period of steep linear slope down in $\mathbf{D}_n$ ($\mathbf{A}_n$), which can correspond to jumps down in the associated limit process. The jump down from $\mathbf{D}_n$ ($\mathbf{A}_n$) corresponds to a jump up (down) in the limit process for $\mathbf{Q}_n$.

6.4.3. Simulation Examples

What we intend to do now is simulate and plot the waiting-time and queue-length processes under various assumptions on the interarrival-time and service-time distributions. Just as with the empirical cdf in Example 1.1.1 and the renewal process in Section 6.3, when we plot the queue-length process we need to plot a portion of a continuous-time process. Just as in the two previous cases, we can plot the queue-length process with the statistical package *S*, exploiting underlying random sequences. Here the relevant underlying random sequences are the arrival times $\{T_k\}$ and the departure times $\{D_k\}$, defined recursively above in (4.1) and (4.2).

Since the plotting procedure is less obvious now, we specify it in detail. We first form two dimensional vectors by appending a +1 to each arrival time and a -1 to each departure time. (Instead of the arrival time T_n , we have the vector $(T_n, 1)$; instead of the departure time D_n , we have the vector $(D_n, -1)$.) We then combine all the vectors (creating a matrix) and sort on the first component. The new first components are thus the successive times of any change in the queue length (arrival or departure). We then form the successive cumulative sums of the second components, which converts the second components into the queue lengths at the times of change. We could just plot the queue lengths at the successive times of change, but we go further to plot the full continuous-time queue-length process. We can plot by linear interpolation, if we include each queue length value twice, at the jump when the value is first attained and just before the next jump. (This method inserts a vertical line at each jump.)

We now give an *S* program to read in the first n interarrival times, service times and waiting times and plot the queue-length process over the time interval that these n customers are in the system (ignoring all subsequent arrivals). At the end of the time interval the system is necessarily empty. Our construction thus gives an odd end effect, but it can be truncated. Indeed, in our plots below we do truncate (at the expected time of the n^{th} arrival).

Here is the *S* function:

```

QueueLength <- function(U, V, W) {
  QueueLength <- vector("numeric", 2*length(U) + 1)
  T <- cumsum(U)                #construct arrival times
  D <- T + W + V                # departure times
  TT <- cbind(T, +1)            #append +1 to arrivals
  DD <- cbind(D, -1)            #append -1 to deps.
  m <- rbind(TT, DD)            #merge into one matrix
  msort <- m[sort.list(m[, 1]),] #sort on first comp.
  time1 <- msort[, c(1)]         #extract change times
  QLchg <- msort[, c(2)]         #queue length changes
  QL1 <- cumsum(QLchg)           #successive q. lths.
  time2 <- c(0, time1, time1)    #times for lin.interp.
  time <- sort(time2)
  n <- length(time1)             #q. lths. for lin. int.
  QL <- c(0, QL1)
  for (k in seq(n)) {
    QueueLength[[2 * k - 1]] <- QL[[k]]
    QueueLength[[2 * k]] <- QL[[k]] }
  QueueLength[2 * n + 1] <- QL[n + 1]
  plot(time, QueueLength, type = "l") #do the plotting
}

```

We now consider a few examples. We use the *Kendall notation* to describe the model: $X/Y/c$ specifies a model with c servers, arrival process of type X and service process of type Y . For either X or Y , GI denotes an IID sequence with a general distribution, while M (for Markov) denotes (in addition) the exponential distribution. We use P_p for the Pareto distribution with parameter p .

Example 6.4.1. *The M/M/1 Queue.*

We first consider the standard M/M/1 queue. Thus, here we assume that the interarrival times and service times come from mutually independent sequences of IID exponentially distributed random variables. It suffices to specify the means of the interarrival time and the service time. Using the scaling in equations (4.9) and (4.10), we need to specify the constant m and the space-scaling sequence $\{c_n : n \geq 1\}$.

At this point, we know what to do: There are no heavy-tailed distributions, so we should let $c_n = \sqrt{n}$. We also let $m = 1$. Thus, we fully specify

the sequence of $M/M/1$ models indexed by n by letting

$$EU_k^n = 1 + 1/\sqrt{n} \quad \text{and} \quad EV_k^n = 1 \quad \text{for all } k \quad \text{and } n. \quad (4.19)$$

With that choice, the plotter can do the appropriate scaling automatically.

We are primarily interested in the queue-length process, but we also plot the waiting times, because it is instructive to compare the plotted queue-length process to the plotted waiting times. Hence, we plot both the waiting times of the first n customers (linearly interpolated) and the queue-length process over the time interval $[0, nEU_1^n]$ for the cases $n = 10^j$ with $j = 1, 2, 3$ in Figure 6.8.

For small n , the queue-length process looks very different from the waiting time sequence, but as n increases, the sample path of the queue length process becomes very similar to the sample path of the waiting times, except possibly for the final portion, where the queue length experiences some of the end effect. To confirm what we see in Figure 6.8, we plot three possible realizations of the waiting times and the queue lengths for $n = 10^4$ in Figure 6.9.

From our experience so far, we should know what to expect: The plots are approaching plots of reflected Brownian motion with drift -1 (which does not have any jumps). Now the conditions and conclusions of Theorem 6.4.1 hold with $c_n = \sqrt{n}$ and $\mathbf{W} = \phi(\sigma\mathbf{B} - m\mathbf{e})$, where $\mathbf{B}$ is standard Brownian motion, $\mathbf{e}$ is the identity map, $\phi : D \rightarrow D$ is the one-sided reflection map and $\sigma^2 = \text{Var}(U_1) + \text{Var}(V_1) = 2$. We apply Donsker's theorem – Theorem 4.3.2.

Moreover, the plots show that the distance between the two scaled processes is indeed asymptotically negligible. Since the limit process here has continuous sample paths, we can express this asymptotic equivalence using the uniform norm over $[0, 1]$:

$$\|\mathbf{W}_n - \mathbf{Q}_n\| \Rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (4.20)$$

■

Example 6.4.2. *The $M/P_{1.5}/1$ Queue.*

We now modify the previous example by letting the service-time distribution be Pareto(p) with $p = 1.5$ and mean 1. (In the framework of Section 1.3.3, we can use $3^{-1}U^{-2/3}$, where U is uniform on the interval $[0, 1]$, which has cdf $F^c(t) = (3t)^{-3/2}$ for $t \geq 1$.) With this heavy-tailed service-time distribution, we must scale space differently, because the space scaling in the

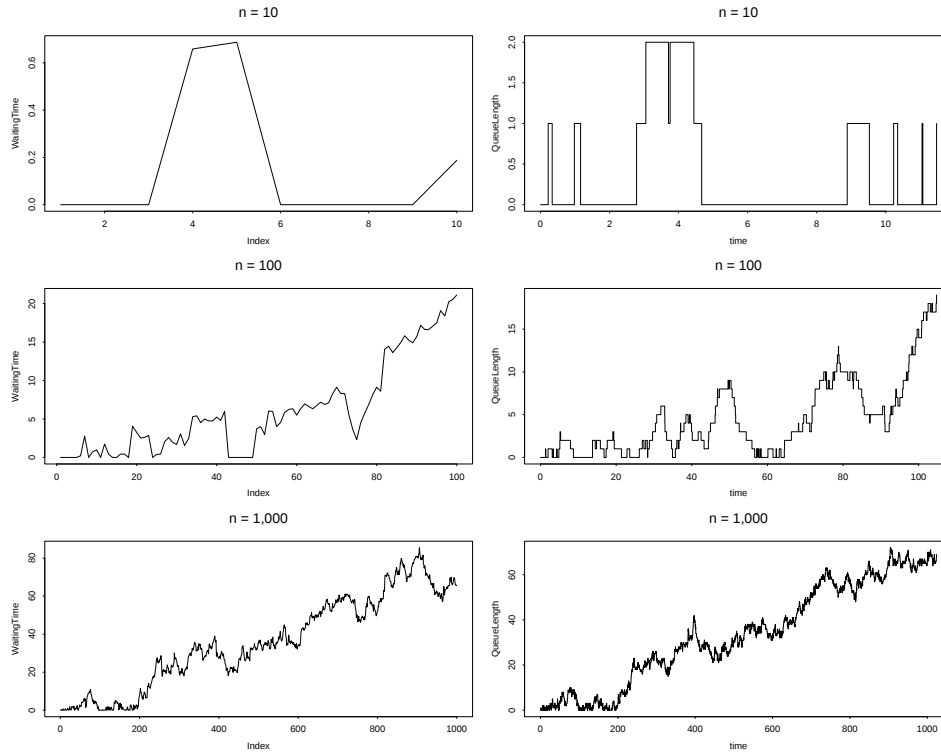


Figure 6.8: Plots of the waiting times of the first n arrivals (on the left) and the queue-length process over the interval $[0, nEU_1^n]$ (on the right) in the M/M/1 queue with scaling in (4.19) for $n = 10^j$ with $j = 1, 2, 3$.

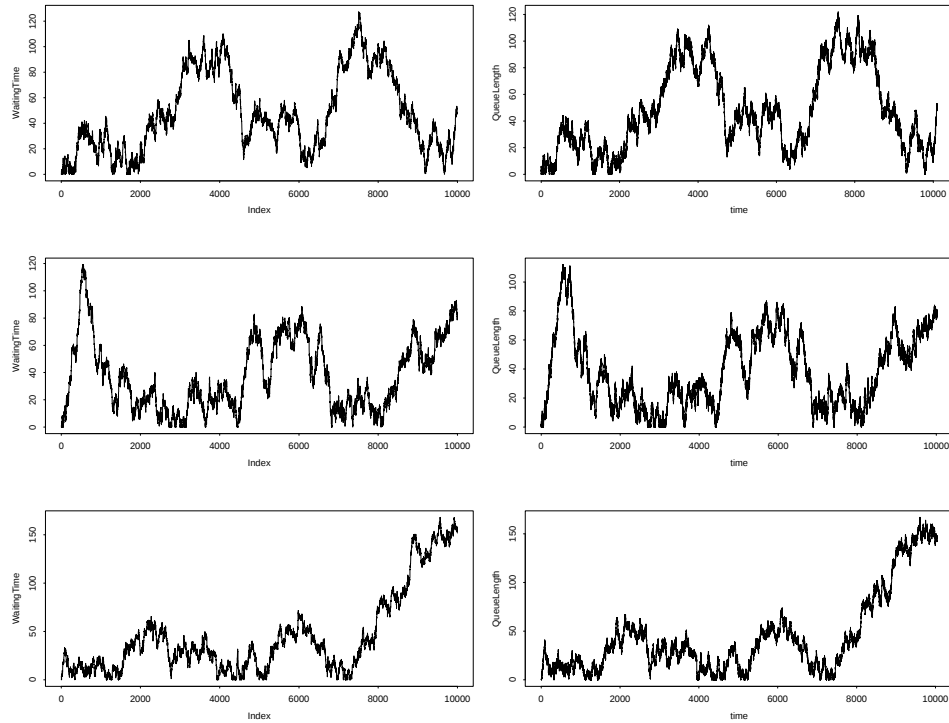


Figure 6.9: Three possible realizations of the waiting times of the first n arrivals (on the left) and the queue-length process over the interval $[0, nEU_1^n]$ (on the right) in the M/M/1 queue with scaling in (4.19) for $n = 10^4$.

FCLT for the random walk involves $c_n = n^{2/3}$ instead of $c_n = n^{1/2}$. Hence, instead of the scaling in (4.19), we now use

$$EU_k^n = 1 + n^{-1/3} \quad \text{and} \quad EV_k^n = 1 \quad \text{for all } k \text{ and } n. \quad (4.21)$$

The new scaling makes the traffic intensity ρ_n smaller than in Example 6.4.1 for any given n . For example, for $n = 10,000$, before we had $\rho_n = 1/1.01 \approx 0.990$, while now we have $\rho_n \approx 1/1.046 \approx 0.956$.

We plot three possible realizations of the waiting times of the first n customers (on the bottom or left) and queue-length process over the interval $[0, nEU_k^n]$ (on the top or right) for $n = 10^4$, in Figure 6.10. The first two plots look much like the $M/M/1$ plots in Figure 6.9 except now we can see upward jumps. But the third plot is very different!

There is now much more variability in the sample paths because of the possibility of the occasional very large jumps. The range of values is exceptionally small in case 2 and exceptionally large in case 3. The possibility of exceptionally large jumps produces large variations from plot to plot, as we saw for the random walks in Figure 1.21.

When we look at the third plots closely, it is not evident that the waiting-time and queue-length plots are for the same sample path. For instance, the second big jump in the waiting times occurs at about index 3100, whereas the corresponding second steep incline in the queue-length path begins at about time 4100. However, upon reflection, we see that these actually are consistent, because the waiting time of the customer having the second large service time is about 1000. Since the arrival rate is 1, that customer arrives at about time 3100. Hence that customer enters service, and begins occupying the server, at about time 4100. Thus the queue length should start building up at about time 4100, as it does.

The upward jumps are less sharp for the queue-length process, which we know actually increases by unit jumps, but the asymptotic behavior is evident from the plots. In this case, we are seeing a reflected stable Lévy motion with drift -1 , which has discontinuous sample paths, instead of a reflected Brownian motion. Again we can explain the statistical regularity we see by Theorem 6.4.1. However, now the scaling involves $c_n = n^{2/3}$.

By Theorems 4.5.2 and 4.5.3, the limit process is $\mathbf{W} \equiv \phi(\sigma \mathbf{S}^v - \mathbf{e}) \equiv \sigma \phi(\mathbf{S}^v - \sigma^{-1} \mathbf{e})$, where $\sigma = 1/3C_\alpha^{2/3}$ for C_α in (5.14) of Section 4.5.1, $\mathbf{S}^v$ is a centered α -stable Lévy motion with $\mathbf{S}^v(1) \stackrel{d}{=} S_\alpha(1, 1, 0)$ and $\alpha = 3/2$. (Its steady-state distribution is given in Section 8.5.2.) Again, it is evident that the two scaled processes $\mathbf{W}_n$ and $\mathbf{Q}_n$ should now be asymptotically equivalent. ■

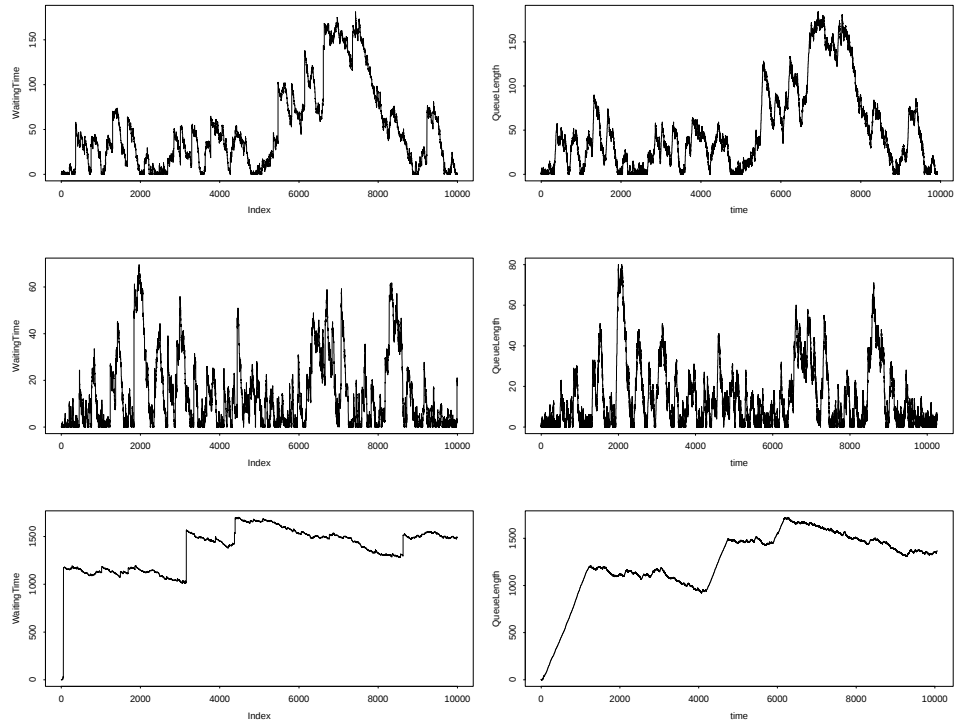


Figure 6.10: Three possible realizations of the waiting times of the first n arrivals (on the left) and the queue-length process over the interval $[0, nEU_1^n]$ (on the right) in the $M/P_{1.5}/1$ queue with the scaling in (4.21) for $n = 10^4$.

Example 6.4.3. *The $P_{1.5}/M/1$ Queue.*

It is evident that a heavy-tailed service-time distribution should cause greater congestion, but it may not be evident that a heavy-tailed interarrival-time distribution can as well, because extra long interarrival times only serve to empty out the queue. However, heavy-tailed interarrival-time distributions can cause congestion as well. The reason is that, for given fixed mean, the occasionally exceptionally long interarrival times must be compensated for in the distribution by shorter interarrival times, and these shorter interarrival times lead to bursts of arrivals and thus increased queue lengths.

We illustrate by considering the $P_{1.5}/M/1$ queue, which has IID Pareto(1.5) interarrival times and IID exponential service times. This model is the *dual* of the model in Example 6.4.2, with the role of the interarrival times and service times switched (adjusted by scaling, so that the expected interarrival times are bigger than the expected service times in both cases).

In Figure 6.11 we plot three possible realizations of the waiting times of the first n arrivals (on the left) and the queue-length process over the interval $[0, nEU_1^n]$ (on the right) in the $P_{1.5}/M/1$ queue with the scaling in (4.21) for $n = 10^4$.

As in Figures 6.8 – 6.10, the queue-length plots are similar to the waiting-time plot, except possibly for the final portion of the queue-length plot, where the queue experiences its end effect. However, unlike in the previous figures, in Figure 6.11 we see evidence of jumps down.

Just as for the $M/P_{1.5}/1$ model, the heavy-traffic FCLT in Theorem 6.4.1 applies to the $P_{1.5}/M/1$ and $P_{1.5}/P_{1.5}/1$ models. Indeed, we again have the same scaling, but now the limiting reflected stable Lévy motions are different, having jumps down only for the $P_{1.5}/M/1$ model and having jumps both up and down for the $P_{1.5}/P_{1.5}/1$ model, instead of having jumps up only for the $M/P_{1.5}/1$ model.

For the $P_{1.5}/M/1$ model, the heavy-traffic stochastic-process limit for the workload process is $\mathbf{W}_n \Rightarrow \mathbf{W}$, where again $c_n = n^{2/3}$, but now

$$\mathbf{W} = \phi(-\sigma \mathbf{S}^u - \mathbf{e}) \stackrel{d}{=} \sigma \phi(-\mathbf{S}^u - \sigma^{-1} \mathbf{e}) ,$$

where $\sigma = 1/3C_\alpha^{2/3}$ for $\alpha = 3/2$, just as in Example 6.4.2. Here $-\mathbf{S}^u(1) \stackrel{d}{=} S_\alpha(1, -1, 0)$.

For the $P_{1.5}/P_{1.5}/1$ model, the limit process is

$$\mathbf{W} = \phi(\sigma \mathbf{S}^v - \sigma \mathbf{S}^u - \mathbf{e}) \stackrel{d}{=} \sigma \phi(\mathbf{S}^v - \mathbf{S}^u - \sigma^{-1} \mathbf{e}) ,$$

where $\mathbf{S}^v - \mathbf{S}^u \stackrel{d}{=} \mathbf{S}$ with $\mathbf{S}$ being a stable Lévy motion satisfying $\mathbf{S}(1) \stackrel{d}{=} 2^{2/3} S_\alpha(1, 0, 0)$; see (5.8) – (5.11) in Section 4.5.1.

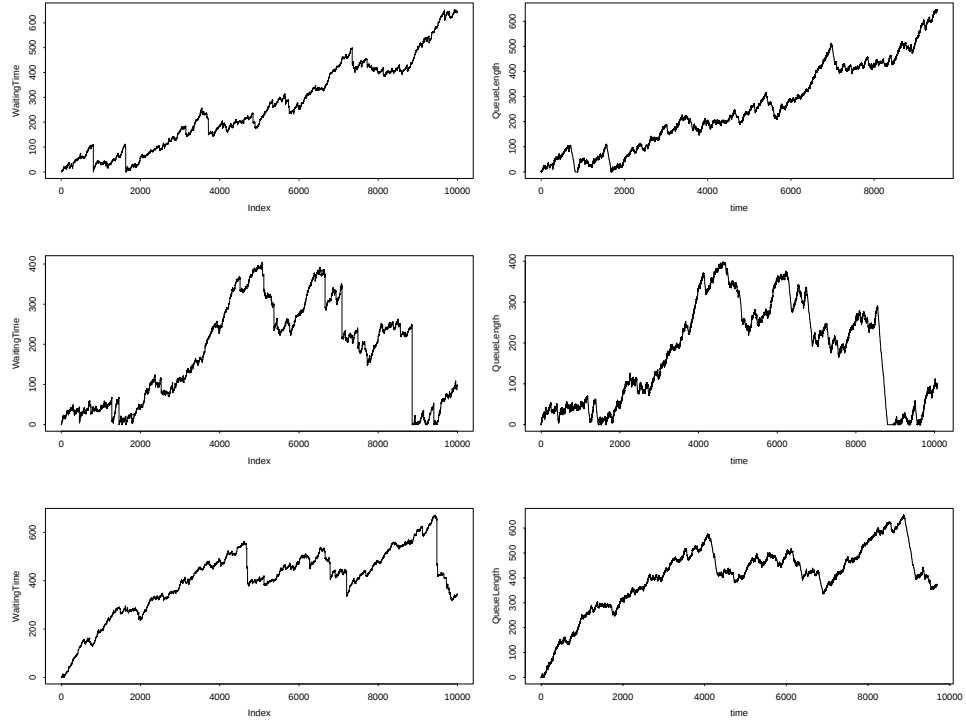


Figure 6.11: Three possible realizations of the waiting times of the first $n = 10^4$ arrivals (on the left) and the queue-length process over the interval $[0, nEU_1^n]$ (on the right) in the $P_{1.5}/M/1$ queue with the scaling in (4.21) for $n = 10^4$.

We should not be fooled by the jumps down for the $P_{1.5}/M/1$ model. Of course, the jumps down do constitute reductions in congestion, but elsewhere in the plot the sample path is rising, so that the range of values experienced can be substantial. Indeed, that is demonstrated by the heavy-traffic FCLT, which has space scaling by $n^{2/3}$, just as for the $M/P_{1.5}/1$ model in Example 6.4.2. ■

6.5. Rare Long Service Interruptions

The queueing example just considered illustrates a common cause of congestion in queues: stochastic variability in the interarrival times and service times. However, congestion in queues can occur for other reasons: For example, the servers may be subject to breakdown and failure, causing service interruptions. In manufacturing systems, service interruptions due to machine failures or the unavailability of parts are often the dominant sources of congestion. With evolving communication networks, there is debate about whether the most important source of congestion is the uncertain burstiness of customer input or the uncertain failure of system elements. The biggest problems tend to occur when both happen together.

We can better understand the impact of service interruptions upon performance if we develop a probability model and establish appropriate stochastic-process limits. One such model, considered by Kella and Whitt (1990), is a queue with rare long service interruptions. The queue can be a standard single-server queue with unlimited waiting space, the first-come first-served service discipline and random arrivals and service times, as considered in the previous section. We can supplement that model by allowing random service interruptions. The interruptions can be triggered by queueing events; e.g., they could occur only when the queue becomes empty. Or they can occur exogenously. We will consider the case in which they occur exogenously.

Specifically, we will assume that the availability of the server is characterized by an alternating renewal process; i.e., there are alternating periods in which the server is available (up) or unavailable (down). For tractability, we assume that the up and down times come from mutually independent sequences of IID positive random variables with finite means and variances.

A revealing stochastic-process limit can be obtained by considering the queue in a heavy-traffic limit, in which the load is allowed to approach the critical value for stability. If the interruptions remain unchanged, then the service interruptions alter the conventional heavy-traffic limit with a reflected Brownian motion limit process only by increasing the traffic intensity and increasing the variance parameter of the Brownian motion, both of

which cause increased congestion. However, we obtain a different nondegenerate limit, which is consistent with many applications, if we let the intervals between interruptions and the durations of the interruptions increase in the limit. If we let these quantities increase appropriately, with the duration of an interruption being asymptotically negligible compared to the time between interruptions, then we can obtain a revealing nondegenerate limit.

In particular, an interesting limiting regime has the random up times be of order n and the random down times be of order $\sqrt{n}$ as a function of the number n of customers being considered. Then, with the customary scaling of time by n and space by $\sqrt{n}$, the scaled up times become of order 1 and the scaled down times become of order $1/\sqrt{n}$. That makes the scaled down times asymptotically negligible. Thus, after scaling, the service interruptions occur in the limit according to a stochastic point process, with a finite positive expected number of interruptions in a finite time interval.

Since the scaled durations of the service interruptions are asymptotically negligible, the service interruptions occur instantaneously in the limit. Nevertheless, the service interruptions can have a significant spatial impact, because the number of arrivals during the order $\sqrt{n}$ down time is also of order $\sqrt{n}$. Thus, after scaling space by $\sqrt{n}$, the input during the down time causes a random jump of order 1 in the scaled queueing process at each interruption time.

The proposed scaling, with up times of order n and down times of order $\sqrt{n}$, thus produces random jumps of order-1 size, spaced at random order-1 intervals. In the limit, the proportion of time that the server is unavailable because of interruption is asymptotically negligible. Nevertheless, the asymptotic impact of the interruptions can be dramatic. With this limit, it is possible to compare the effects of the service interruptions (which appear in the limit process as jumps) to the customary stochastic fluctuations. Depending on the specific parameter settings, one or the other may dominate. In Section 14.7, following Kella and Whitt (1990) and Chen and Whitt (1993), we consider networks of queues with rare long service interruptions.

When we consider limits for sequences of queue-length stochastic processes affected by rare long interruptions of the kind just described, the jumps in the limit process are typically not matched in the converging scaled queue-length processes. In the queueing system, arrivals usually are coming one at a time. During a service interruption, service stops, but the arrivals keep coming. Thus the queue length process increases by many unit steps during such periods. After scaling time and space, the n^{th} scaled queue-length process increases more rapidly (due to the time scaling) but by smaller asymptotically negligible amounts (due to the space scaling). Thus

the resulting limit is a stochastic-process limit with unmatched jumps in the limit process.

In the rest of this subsection we illustrate the kind of limiting behavior provided by rare long service interruptions. To do so, we simplify the model: Even though service interruptions represent a different source of congestion than variability in customer demand, we often can represent service interruptions within the framework of a standard queueing model. We can simply include the interruption in the service time of one of the customers. Specifically, we can redefine the service-time distribution: The new service-time distribution becomes a mixture: With probability p , the new service time is the sum of an original service time and the interruption duration; with probability $1 - p$, the new service time reduces to an original service time. We then choose the probability p to match the probability that a customer is the first customer to experience a service interruption. If the timing of service interruptions needs to be modeled very precisely, then we can think of interruptions as special high-priority customers that preempt regular customers (in line or in service), but the simple model above often suffices.

We have in mind rare long service interruptions occurring randomly, but to illustrate the interruption phenomenon, we let the interruptions occur in a fixed manner in our example below.

Example 6.5.1. *The $M/M/1$ queue with two fixed service interruptions.*

We construct a simple example to illustrate the kind of limit behavior associated with rare long service interruptions. Specifically, we consider the $M/M/1$ queue with the heavy-traffic scaling in (4.19), just as in Example 6.4.1, except that now we let customers number $n/4$ and $3n/4$ have service times of $2\sqrt{n}$ and $\sqrt{n}$, respectively, as a function of n . These special service times are introduced to represent interruptions that occur approximately at times $t/4$ and $3t/4$ in the scaled processes plotted over the interval $[0, 1]$. (By the SLLN, the scaled arrival time of customer number $n/4$ approaches $t/4$ as $n \rightarrow \infty$.) Note that the spacings between the interruptions is indeed order n , while the durations of the interruptions (as captured by the special service times) are of order $\sqrt{n}$, as specified above.

We plot the waiting times of the first n customers and the queue-length process for the time interval $[0, nEU_1^n]$, the expected time for the n customers to arrive, for $n = 10^j$ with $j = 2, 3, 4$ in Figure 6.12. In Figure 6.12 the impact of the interruptions is clearer for the waiting times than for the queue lengths, especially for smaller n . For the queue-length process, the portion of the plot corresponding to the jump gets steeper as n increases.

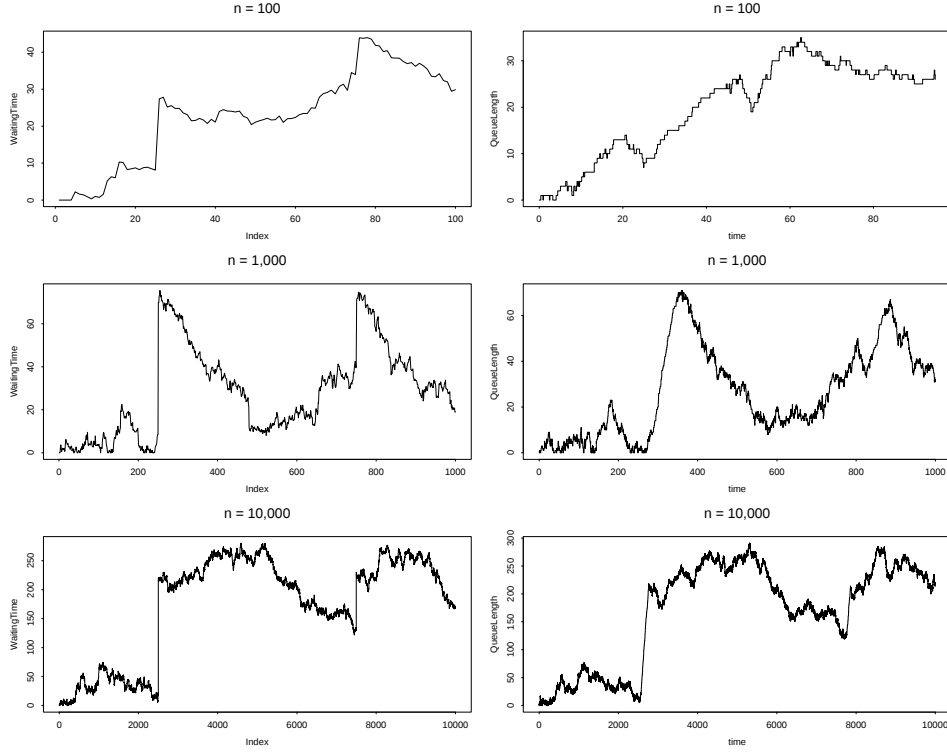


Figure 6.12: Plots of the waiting times of the first n arrivals (on the left) and the queue-length process over the interval $[0, nEU_1^n]$ (on the right) for in the $M/M/1$ queue with scaling in (4.19) and service interruptions of length $2\sqrt{n}$ and $\sqrt{n}$ associated with customers $n/4$ and $3n/4$ for $n = 10^j$ with $j = 2, 3, 4$.

As before, we see that the queue-length and waiting-time plots coalesce as n increases. Now both scaled processes approach reflected Brownian motion with drift -1 , modified by jumps of size 2 at time $t = 1/4$ and of size 1 at time $t = 3/4$. For the scaled queue-length process, the limit process must have unmatched jumps. ■

Example 6.5.2. *The $P_{1.5}/M/1$ queue with two fixed service interruptions.*

Now, as in Example 6.4.3 we consider the $P_{1.5}/M/1$ queue with heavy-traffic scaling in (4.21), modified by having customers number $n/4$ and $3n/4$ experience interruptions. We choose the $P_{1.5}/M/1$ model instead of the $M/P_{1.5}/1$ model, because it naturally (without the interruptions) produces jumps down instead of up. Thus, it will be easier to recognize the new jumps up caused by the service interruptions.

In addition, the durations of the interruptions need to be scaled differently from the scaling in Example 6.5.1. In order to be consistent with the heavy-traffic limiting behavior in Example 6.4.3, we now need to scale the durations of the interruptions by $n^{2/3}$ instead of $n^{1/2}$. In particular, now we let the service times of customers number $n/4$ and $3n/4$ be $2n^{2/3}$ and $n^{2/3}$, respectively. We plot three possible realizations of the waiting times of the first n customers and the queue-length process over the time interval $[0, nEU_1^n]$, ignoring all arrivals after the first n , for the case $n = 10^4$ in Figure 6.13.

Just as we would expect from Figures 6.11 and 6.12, we see randomly occurring jumps down because of the $P_{1.5}$ arrival process and jumps up of magnitude 2 at time $t = 1/4$ and 1 at time $t = 3/4$. However, both kinds of jumps are much sharper for the waiting times than for the queue-length process. Hence, we evidently need larger n in this case to have the queue-length plots be visually similar to the waiting-time plots. The supporting FCLTs state that both scaled processes converge to a stable Lévy motion (with jumps down only) modified by the addition of two jumps up, a jump of size 2 at $t = 1/4$ and a jump of size 1 at $t = 3/4$; again, see Sections 4.5 and 14.7. Again, for the scaled queue-length process, that limit process must have unmatched jumps. ■

The simple models of service interruptions considered in Examples 6.5.1 and 6.5.2 are of course quite artificial. However, from these examples, we can anticipate what we will see when we use the more realistic alternating renewal process model for up and down times.

6.6. Time-Dependent Arrival Rates

In many service systems, congestion occurs primarily because of systematic, deterministic variations in the input rate over time. Many service systems have arrival rates that vary systematically with time, so that there are known busy periods with higher loads than average. However, everything is not known. There remains uncertainty about the actual input; there are unanticipated fluctuations about the known time-varying deterministic rates.

To better understand the behavior of queues with time-varying arrival rates, we need to focus directly on queueing models with time-varying arrival rates. Just as for stationary queueing models, it can be helpful to consider heavy-traffic limits for queues with time-varying arrival rates. With time-varying arrival rates, we still scale time, but we think of expanding time

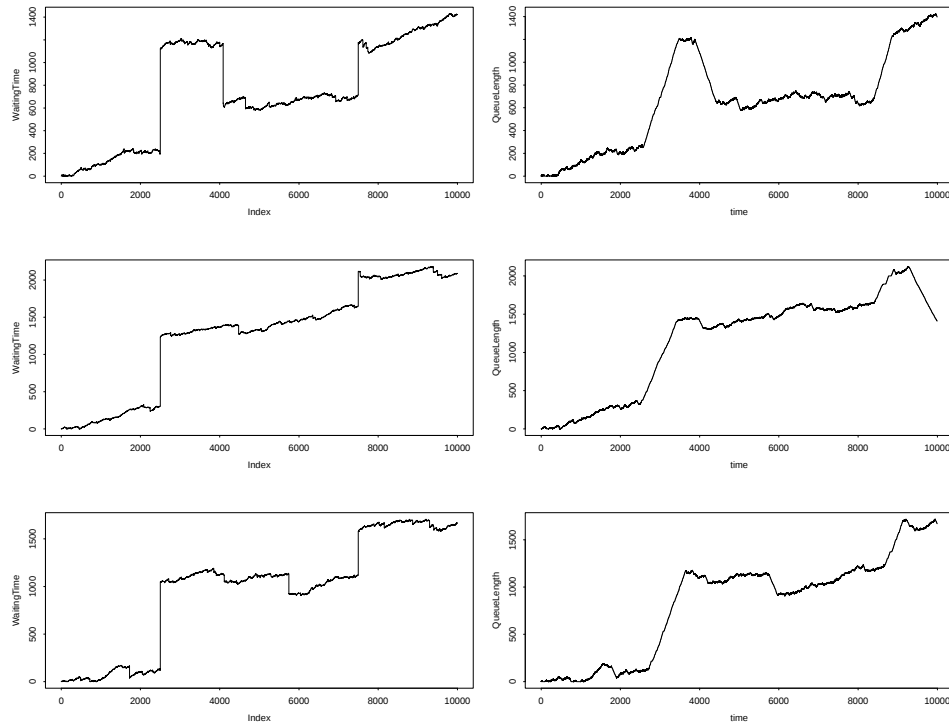


Figure 6.13: Three possible realizations of the waiting times of the first n arrivals (on the left) and the queue-length process over the interval $[0, nEU_1^n]$ (on the right) in the $P_{1.5}/M/1$ queue with scaling in (4.21) and service interruptions of length $2n^{2/3}$ and $n^{2/3}$ associated with customers $n/4$ and $3n/4$ for $n = 10^4$.

immediately prior to the time of interest. We increase the overall arrival and service rate, which is tantamount to decreasing the rate of change in the arrival-rate and service-rate functions, so that temporary periods of overload or underload before the time of interest tend to persist longer and longer.

With such scaling, a law of large numbers can be established, in which the scaled queue-length process converges to a reflection of a deterministic net-input process, where the limiting deterministic net-input process satisfies an *ordinary differential equation* (ODE) driven by the original time-dependent arrival and service rates. That limit is identical to the direct deterministic ODE approximation we obtain if we ignore the stochastic aspects of the model. In the direct deterministic approximation, the net input becomes the solution an ODE driven by the time-dependent arrival and service rates; i.e., if λ is the arrival-rate function and μ is the service-rate function, then the deterministic approximation for the queue length is the function q satisfying

$$q(t) = \phi(x)(t) \equiv x(t) - \inf_{0 \leq s \leq t} x(s), \quad t \geq 0, \quad (6.1)$$

where ϕ is again the one-sided reflection map, $q(0)$ is the initial queue length (assumed to satisfy $q(0) = 0$) and x is the deterministic net-input function, satisfying the ODE

$$\dot{x}(t) = \lambda(t) - \mu(t), \quad t \geq 0. \quad (6.2)$$

When the deterministic fluctuations dominate the stochastic fluctuations, such a deterministic analysis can be very useful to describe system performance; e.g., see Oliver and Samuel (1962), Newell (1982) and Hall (1991).

However, in stochastic-process limits, we are primarily interested in going beyond the deterministic ODE limit described above. For example, Mandelbaum and Massey (1995) show that it is possible to establish a stochastic (FCLT) refinement to the deterministic ODE limit. It again can be obtained by applying the continuous-mapping approach to stochastic-process limits. In this setting, the continuous-mapping approach involves convergence preservation with nonlinear centering, and can be approached by identifying the directional derivative of the reflection map; see Chapter 6 of the Internet Supplement.

The behavior of the limit process in the stochastic-process limit depends on the deterministic function q . At any time, the deterministic function q must be in one of three states (based on the history of the build up prior to the time of interest): overloaded, critically loaded (when the cumulative input rate is in balance with the output rate) or underloaded. (Roughly

speaking, these regimes correspond to the three cases $\rho > 1$, $\rho = 1$ and $\rho < 1$ in a stationary queueing model.)

With the usual stochastic assumptions (without any heavy-tailed distributions), the stochastic-process refinement is a diffusion process centered about the deterministic function q . The diffusion process corresponds to: ordinary Brownian motion when q is overloaded, reflected Brownian motion when q is critically loaded, and the zero function when q is underloaded.

Within each region, i.e., within any interval in which the deterministic function q remains in one of its three basic states (overloaded, critically loaded or underloaded), the limiting stochastic process has continuous sample paths, but at the boundaries between different regions the limiting stochastic process can have jumps that are unmatched in the converging processes. Thus, the boundary points between different regions for the deterministic function q act as phase transitions for the queueing system. Relatively abrupt changes in the queueing process can occur at these transition times. And, once again, we have a stochastic-process limit with unmatched jumps.

Example 6.6.1. *A shift from critically loaded to underloaded.*

We now give a simple example. In the standard situation we have in mind, the arrival-rate function is changing continuously, so that we can obtain the deterministic net-input function by solving the ODE in (6.2). However, now we consider the more elementary situation in which there is a sudden shift down in the arrival rate at one time. As in the standard situation, we let the service rate be constant (although that is not required).

We let the queue initially be critically loaded, i.e., with $\rho = 1$, and then in the middle of the time period, we reduce the arrival rate, making the model underloaded. For simplicity, we again use the $M/M/1$ queue. We let the mean service time always be 1. We actually deviate slightly from the prescription for the arrival rate: We let the mean interarrival time for the first $n/2$ customers be 1 and the mean interarrival time of the next $n/2$ customers be 2. Hence, after $n/2$ arrivals, the instantaneous traffic intensity suddenly shifts from $\rho = 1$ to $\rho = 0.5$. Of course, with this definition, the shift in arrival rate occurs at a random time instead of a deterministic time, but after scaling time by n , that scaled random shift time converges to $t/2$ w.p.1. Thus, what we do is essentially the same as if we let the arrival-rate shift occur exactly at time $n/2$ when we consider n arrivals.

For the specified model, we plot the waiting times of the first n customers and the queue-length process over the time interval $[0, n]$ for $n = 10^j$ for $j = 2, 3, 4$ in Figure 6.14. As in previous plots, the situation is somewhat

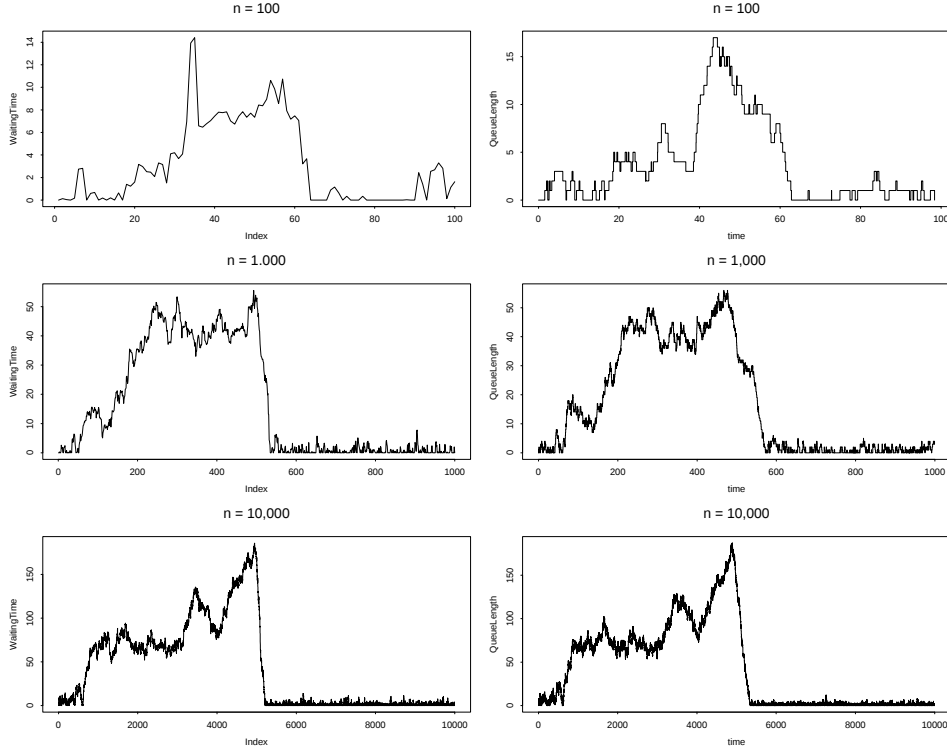


Figure 6.14: Plots of the waiting times of the first n arrivals (on the left) and the queue-length process over the interval $[0, n]$ (on the right) in the $M/M/1$ queue with $\rho = 1$ for the first $n/2$ arrivals and $\rho = 1/2$ for the last $n/2$ arrivals for $n = 10^j$ with $j = 2, 3, 4$.

ambiguous for smaller n , but as n increases, we see statistical regularity. As before, the scaled waiting-time and queue-length plots coalesce as n increases. As n increases, a sharp jump down is visible when the traffic intensity shifts from $\rho = 1$ to $\rho = 1/2$. As we indicated before, asymptotically, this shift for the scaled processes occurs at time $t = 1/2$.

Again, we are able to establish supporting FCLTs. Both the scaled waiting-time process and the scaled queue-length process are approaching reflecting Brownian motion over the subinterval $[0, t/2)$ and the 0 function over the subinterval $[t/2, 1]$. As in the previous examples, the scaled queue-length and waiting-time processes are asymptotically equivalent.

Thus, the limit process for the scaled queue-length process has an unmatched jump at $t = 1/2$. In this example, the limit for the waiting-time

process also has an unmatched jump at the same time.

Chapter 12

The Space D

12.1. Introduction

This chapter is devoted to the function space $D \equiv D([0, T], \mathbb{R}^k)$ with the Skorohod M_1 topology, expanding upon the introduction in Sections 3.3 and 11.5 and the classic paper by Skorohod (1956). We omit most proofs here. Many are provided in Chapter 6 of the Internet Supplement.

Here is how the present chapter is organized: We start in Section 12.2 by discussing regularity properties of the function space D . A key property, which we frequently use, is the fact that any function in D can be approximated uniformly closely by piecewise-constant functions with only finitely many discontinuities.

In Section 12.3 we introduce the strong and weak versions of the M_1 topology on $D([0, T], \mathbb{R}^k)$, referred to as SM_1 and WM_1 , and establish basic properties. We also discuss the relation among the non-uniform Skorohod topologies on D . In Section 12.4 we discuss local uniform convergence at continuity points and relate it to oscillation functions used to characterize different forms of convergence.

In Section 12.5 we provide several different alternative characterizations of SM_1 and WM_1 convergence. Some involve parametric representations of the completed graphs and others involve oscillation functions. It is significant that there are forms of the oscillation-function characterizations that involve considering one function argument t at a time. Consequently, the examples in Figure 11.2 tend to be more than illustrative: The topologies are characterized by the local behavior in the neighborhood of single discontinuities.

In Section 12.6 we discuss conditions that allow us to strengthen the mode of convergence from WM_1 to SM_1 . The key condition is to have the

coordinate limit functions have no common discontinuities. In Section 12.7 we study how SM_1 convergence in $D([0, T], \mathbb{R}^k)$ can be characterized by associated limits of mappings.

In Section 12.8 we exhibit a complete metric topologically equivalent to the incomplete metric inducing the SM_1 topology introduced earlier. As with the J_1 metric d_{J_1} in (3.2) of Section 3.3, the natural M_1 metric is incomplete, but there exists a topologically equivalent complete metric, so that D with the SM_1 topology is Polish (metrizable as a complete separable metric space).

In Section 12.9 we discuss extensions of the SM_1 and WM_1 topologies on $D([0, T], \mathbb{R}^K)$ to corresponding spaces of functions with non-compact domains. The principal example of such a non-compact domain is the interval $[0, \infty)$, but $(0, \infty)$ and $(-\infty, \infty)$ also arise.

In Section 12.10 we introduce the strong and weak versions of the M_2 topology, denoted by SM_2 and WM_2 . In Section 12.11 we provide alternative characterizations of these topologies and discuss additional properties.

Finally, in Section 12.12 we discuss characterizations of compact subsets of D using oscillation functions. These characterizations are useful because they lead to characterizations of tightness for sequences of probability measures on D , which is a principal way to establish weak convergence of the probability measures; see Section 11.6.

12.2. Regularity Properties of D

Let $D \equiv D^k \equiv D([0, T], \mathbb{R}^k)$ be the set of all $\mathbb{R}^k$ -valued functions $x \equiv (x^1, \dots, x^k)$ on $[0, T]$ that are right continuous at all $t \in [0, T)$ and have left limits at all $t \in (0, T]$: If $x \in D$, then

$$\text{for } 0 \leq t < T, \quad x(t+) \equiv \lim_{s \downarrow t} x(s) \quad \text{exists with} \quad x(t+) = x(t)$$

and

$$\text{for } 0 < t \leq T, \quad x(t-) \equiv \lim_{s \uparrow t} x(s) \quad \text{exists}.$$

However, with the M_1 topology, we will be working with the completed graphs of the functions, which are obtained by adding segments joining the left and right limits to the graph at each discontinuity point. Thus the actual value of the function at discontinuity points does not matter, provided that the function value falls appropriately between the left and right limits. Such functions are said to have *discontinuities of the first kind*. In Chapter 15 we consider more general functions.

We use superscripts to designate coordinate functions, so that subscripts can index different functions in D . For example, x_3^2 denotes the second coordinate function in $D([0, T], \mathbb{R}^1)$ of $x_3 \equiv (x_3^1, \dots, x_3^k)$ in $D([0, T], \mathbb{R}^k)$, where x_3 is the third element of the sequence $\{x_n : n \geq 1\}$. Let C be the subset of continuous functions in D .

Let $\|\cdot\|$ be the maximum (or l_∞) norm on $\mathbb{R}^k$ and the *uniform norm* on D ; i.e., for each $b \equiv (b^1, \dots, b^k) \in \mathbb{R}^k$, let

$$\|b\| \equiv \max_{1 \leq i \leq k} |b^i| \quad (2.1)$$

and, for each $x \equiv (x^1, \dots, x^k) \in D([0, T], \mathbb{R}^k)$, let

$$\|x\| \equiv \sup_{0 \leq t \leq T} \|x(t)\| = \sup_{0 \leq t \leq T} \max_{1 \leq i \leq k} |x^i(t)|. \quad (2.2)$$

The maximum norm on $\mathbb{R}^k$ in (2.1) is topologically equivalent to the l_p norm

$$\|b\|_p \equiv \left(\sum_{i=1}^k (b^i)^p \right)^{1/p}.$$

For $p = 2$, the l_p norm is the Euclidean (or l_2) norm. For $p = 1$, the l_p norm is the sum (or l_1) norm. The uniform norm on D induces the uniform metric on D .

We first discuss regularity properties of D due to the existence of limits. Let $Disc(x)$ be the set of discontinuities of x , i.e.,

$$Disc(x) \equiv \{t \in (0, T] : x(t-) \neq x(t)\} \quad (2.3)$$

and let $Disc(x, \epsilon)$ be the set of discontinuities of magnitude at least ϵ , i.e.,

$$Disc(x, \epsilon) \equiv \{t \in (0, T] : \|x(t-) - x(t)\| \geq \epsilon\}. \quad (2.4)$$

The following is a key regularity property of D .

Theorem 12.2.1. (the number of discontinuities of a given size) *For each $x \in D$ and $\epsilon > 0$, $Disc(x, \epsilon)$ is a finite subset of $[0, T]$.*

Corollary 12.2.1. (the number of discontinuities) *For each $x \in D$, $Disc(x)$ is either finite or countably infinite.*

We say that a function x in D is *piecewise-constant* if there are finitely many time points t_i such that $0 \equiv t_0 < t_1 < \cdots < t_{m-1} \leq t_m \equiv T$ and x is constant on the intervals $[t_{i-1}, t_i)$, $1 \leq i \leq m-1$, and $[t_{m-1}, T]$. Let D_c be the subset of piecewise-constant functions in D . Let $v(x; A)$ be the *modulus of continuity* of the function x over the set A , defined by

$$v(x; A) \equiv \sup_{t_1, t_2 \in A} \{ \|x(t_1) - x(t_2)\| \} \quad (2.5)$$

for $A \subseteq [0, T]$. The following is a second important regularity property of D .

Theorem 12.2.2. (approximation by piecewise-constant functions) *For each $x \in D$ and $\epsilon > 0$, there exists $x_c \in D_c$ such that $\|x - x_c\| < \epsilon$.*

We can deduce other useful consequences from Theorem 12.2.2.

Corollary 12.2.2. (oscillation function property) *For each $x \in D$ and $\epsilon > 0$, there exist finitely many points t_i with $0 \equiv t_0 < t_1 < \cdots < t_{m-1} \leq t_m \equiv T$ such that $v(x, [t_{i-1}, t_i)) < \epsilon$, $1 \leq i \leq m-1$, and $v(x, [t_{m-1}, T]) < \epsilon$.*

Corollary 12.2.3. (boundedness) *Each x in D is bounded, i.e., $\|x\| < \infty$.*

Corollary 12.2.4. (measurability) *Each x in D is a Borel measurable real-valued function on $[0, T]$.*

12.3. Strong and Weak M_1 Topologies

In this section we define strong and weak versions of the M_1 topology on the function space $D([0, T], \mathbb{R}^k)$, denoted by SM_1 and WM_1 . The strong topology agrees with the standard topology introduced by Skorohod (1956). The strong and weak topologies coincide when $k = 1$ but differ for $k > 1$. We will show that the weak topology coincides with the product topology.

We consider functions with domain $[0, T]$, but our results can be applied to non-compact domains such as $[0, \infty)$, if as is customary we understand $x_n \rightarrow x$ as $n \rightarrow \infty$ in $D([0, \infty), \mathbb{R}^k)$ to mean that the restrictions of x_n to $[0, T]$ converge to the restriction of x to $[0, T]$ for all T that are continuity points of x . We discuss $D([0, \infty), \mathbb{R}^k)$ further in Section 12.9.

12.3.1. Definitions

The strong and weak topologies will be based on different notions of a segment in $\mathbb{R}^k$. For $a \equiv (a^1, \dots, a^k)$, $b \equiv (b^1, \dots, b^k) \in \mathbb{R}^k$, let $[a, b]$ be the *standard segment*, i.e.,

$$[a, b] \equiv \{\alpha a + (1 - \alpha)b : 0 \leq \alpha \leq 1\} \quad (3.1)$$

and let $[[a, b]]$ be the *product segment*, i.e.,

$$[[a, b]] \equiv \bigtimes_{i=1}^k [a^i, b^i] \equiv [a^1, b^1] \times \dots \times [a^k, b^k], \quad (3.2)$$

where the one-dimensional segment $[a^i, b^i]$ coincides with the closed interval $[a^i \wedge b^i, a^i \vee b^i]$, with $c \wedge d = \min\{c, d\}$ and $c \vee d = \max\{c, d\}$ for $c, d \in \mathbb{R}$. Note that $[a, b]$ and $[[a, b]]$ are both subsets of $\mathbb{R}^k$. If $a = b$, then $[a, b] = [[a, b]] = \{a\} = \{b\}$; if $a^i \neq b^i$ for one and only one i , then $[a, b] = [[a, b]]$. If $a \neq b$, then $[a, b]$ is always a one-dimensional line in $\mathbb{R}^k$, while $[[a, b]]$ is a j -dimensional subset, where j is the number of coordinates i for which $a^i \neq b^i$. Always, $[a, b] \subseteq [[a, b]]$.

Remark 12.3.1. *More general range spaces.* We may want to consider the space D with a more general range space than $\mathbb{R}^k$. Generalizations of the M topologies are restricted by the linear structure in the definition of segments in (3.1) and (3.2). However, we can extend the M topologies to Banach-space valued functions. We use that extension to treat the workload process in the infinite-server queue in Section 10.3. ■

We now define completed graphs of the functions: For $x \in D$, let the (standard) *thin graph* of x be

$$\Gamma_x \equiv \{(z, t) \in \mathbb{R}^k \times [0, T] : z \in [x(t-), x(t)]\}, \quad (3.3)$$

where $x(0-) \equiv x(0)$ and let the *thick graph* of x be

$$\begin{aligned} G_x &\equiv \{(z, t) \in \mathbb{R}^k \times [0, T] : z \in [[x(t-), x(t)]]\} \\ &= \{(z, t) \in \mathbb{R}^k \times [0, T] : z^i \in [x^i(t-), x^i(t)] \text{ for each } i\} \end{aligned} \quad (3.4)$$

for $1 \leq i \leq k$. Since $[a, b] \subseteq [[a, b]]$ for all $a, b \in \mathbb{R}^k$, $\Gamma_x \subseteq G_x$ for each x .

We now define *order relations* on the graphs Γ_x and G_x . We say that $(z_1, t_1) \leq (z_2, t_2)$ if either (i) $t_1 < t_2$ or (ii) $t_1 = t_2$ and $|x^i(t_1-) - z_1^i| \leq$

$|x^i(t_1-) - z_2^i|$ for all i . The relation $\leq$ induces a total order on Γ_x and a partial order on G_x .

It is also convenient to look at the ranges of the functions. Let the *thin range* of x be the projection of Γ_x onto $\mathbb{R}^k$, i.e.,

$$\rho(\Gamma_x) \equiv \{z \in \mathbb{R}^k : (z, t) \in \Gamma_x \text{ for some } t \in [0, T]\} \quad (3.5)$$

and let the *thick range* of x be the projection of G_x onto $\mathbb{R}^k$, i.e.,

$$\rho(G_x) \equiv \{z \in \mathbb{R}^k : (z, t) \in G_x \text{ for some } t \in [0, T]\} . \quad (3.6)$$

Note that $(z, t) \in \Gamma_x$ (G_x) for some t if and only if $z \in \rho(\Gamma_x)$ ($\rho(G_x)$). Thus a pair (z, t) cannot be in a graph of x if z is not in the corresponding range.

We now define strong (standard) and weak parametric representations based on these two kinds of graphs. A *strong parametric representation* of x is a continuous nondecreasing function (u, r) mapping $[0, 1]$ onto Γ_x . A *weak parametric representation* of x is a continuous nondecreasing function (u, r) mapping $[0, 1]$ into G_x such that $r(0) = 0$, $r(1) = T$ and $u(1) = x(T)$. (For the parametric representation, “nondecreasing” is with respect to the usual order on the domain $[0, 1]$ and the order on the graphs defined above.) Here it is understood that $u \equiv (u^1, \dots, u^k) \in C([0, 1], \mathbb{R}^k)$ is the spatial part of the parametric representation, while $r \in C([0, 1], [0, T])$ is the time (domain) part. Let $\Pi_s(x)$ and $\Pi_w(x)$ be the sets of strong and weak parametric representations of x , respectively. For real-valued functions x , let $\Pi(x) \equiv \Pi_s(x) = \Pi_w(x)$. Note that $(u, r) \in \Pi_w(x)$ if and only if $(u^i, r) \in \Pi(x^i)$ for $1 \leq i \leq k$.

We use the parametric representations to characterize the strong and weak M_1 topologies. As in (2.1) and (2.2), let $\|\cdot\|$ denote the supremum norms in $\mathbb{R}^k$ and D . We use the definition $\|\cdot\|$ in (2.2) also for the $\mathbb{R}^k$ -valued functions u and r on $[0, 1]$.

Now, for any $x_1, x_2 \in D$, let

$$d_s(x_1, x_2) \equiv \inf_{\substack{(u_j, r_j) \in \Pi_s(x_j) \\ j=1,2}} \{\|u_1 - u_2\| \vee \|r_1 - r_2\|\} \quad (3.7)$$

and

$$d_w(x_1, x_2) \equiv \inf_{\substack{(u_j, r_j) \in \Pi_w(x_j) \\ j=1,2}} \{\|u_1 - u_2\| \vee \|r_1 - r_2\|\} . \quad (3.8)$$

Note that $\|u_1 - u_2\| \vee \|r_1 - r_2\|$ can also be written as $\|(u_1, r_1) - (u_2, r_2)\|$, due to definitions (2.1) and (2.2). Of course, when the range is $\mathbb{R}$, $d_s = d_w = d_{M_1}$ for d_{M_1} defined in (3.4) in Section 3.3.

We say that $x_n \rightarrow x$ in D for a sequence or net $\{x_n\}$ in the SM_1 (WM_1) topology if $d_s(x_n, x) \rightarrow 0$ ($d_w(x_n, x) \rightarrow 0$) as $n \rightarrow \infty$. We start with the following basic result.

12.3.2. Metric Properties

Theorem 12.3.1. (metric inducing SM_1) d_s is a metric on D .

Proof. Only the triangle inequality is difficult. By Lemma 12.3.2 below, for any $\epsilon > 0$, a common parametric representation $(u_3, r_3) \in \Pi_s(x_3)$ can be used to obtain

$$\|u_1 - u_3\| \vee \|r_1 - r_3\| < d_s(x_1, x_3) + \epsilon$$

and

$$\|u_2 - u_3\| \vee \|r_2 - r_3\| < d_s(x_2, x_3) + \epsilon$$

for some $(u_1, r_1) \in \Pi_s(x_1)$ and $(u_2, r_2) \in \Pi_s(x_2)$. Hence

$$d_s(x_1, x_2) \leq \|u_1 - u_2\| \vee \|r_1 - r_2\| \leq d_s(x_1, x_3) + d_s(x_3, x_2) + 2\epsilon .$$

Since ϵ was arbitrary, the proof is complete. ■

To prove Theorem 12.3.1, we use finite approximations to the graphs Γ_x . We first define an order-consistent distance between a graph and a finite subset. We use the notion of a finite ordered subset.

Definition 12.3.1. (order-consistent distance) For $x \in D$, let A be a finite ordered subset of the ordered graph $(\Gamma_x, \leq)$, i.e., for some $m \geq 1$, A contains $m + 1$ points (z_i, t_i) from Γ_x such that

$$(x(0), 0) \equiv (z_0, t_0) \leq (z_1, t_1) \leq \cdots \leq (z_m, t_m) \equiv (x(T), T) . \quad (3.9)$$

The order-consistent distance between A and Γ_x is

$$\hat{d}(A, \Gamma_x) \equiv \sup\{\|(z, t) - (z_i, t_i)\| \vee \|(z, t) - (z_{i+1}, t_{i+1})\|\} , \quad (3.10)$$

where the supremum is over all $(z_i, t_i) \in A$, $1 \leq i \leq m-1$, and all $(z, t) \in \Gamma_x$ such that

$$(z_i, t_i) \leq (z, t) < (z_{i+1}, t_{i+1}) ,$$

using the order on the graph. ■

We now observe that finite ordered subsets A can be chosen to make $\hat{d}(A, \Gamma_x)$ arbitrarily small. The missing proofs are in the Internet Supplement.

Lemma 12.3.1. (finite approximations to graphs) *For any $x \in D$ and $\epsilon > 0$, there exists a finite ordered subset A of Γ_x such that $\hat{d}(A, \Gamma_x) < \epsilon$ for $\hat{d}$ in (3.10).*

To complete the proof of Theorem 12.3.1, we need the following result, which we prove by applying Lemma 12.3.1.

Lemma 12.3.2. (flexibility in choice of parametric representations) *For any $x_1, x_2 \in D$, $(u_1, r_1) \in \Pi_s(x_1)$ and $\epsilon > 0$, it is possible to find $(u_2, r_2) \in \Pi_s(x_2)$ such that*

$$\|u_1 - u_2\| \vee \|r_1 - r_2\| \leq d_s(x_1, x_2) + \epsilon.$$

We will show that the metric d_s induces the standard M_1 topology defined by Skorohod (1956); see Theorem 12.5.1. Since $\Pi_s(x) \subseteq \Pi_w(x)$ for all x , we have $d_w(x_1, x_2) \leq d_s(x_1, x_2)$ for all x_1, x_2 , so that the WM_1 topology is indeed weaker than the SM_1 topology. However, we show below in Example 12.3.2 that d_w in (3.8) is *not* a metric when $k > 1$.

For $x_1, x_2 \in D([0, T], \mathbb{R}^k)$, let d_p be a metric inducing the product topology, defined by

$$d_p(x_1, x_2) \equiv \max_{1 \leq i \leq k} d(x_1^i, x_2^i) \quad (3.11)$$

for $x_j \equiv (x_j^1, \dots, x_j^k)$ and $j = 1, 2$. (Note that $d_s = d_w = d_p$ when the functions are real valued, in which case we use the notation d .) It is an easy consequence of (3.8), (3.11) and the second representation in (3.4) that the WM_1 topology is stronger than the product topology, i.e., $d_p(x_1, x_2) \leq d_w(x_1, x_2)$ for all $x_1, x_2 \in D$. In Section 12.5 we will show that actually the WM_1 and product topologies coincide.

We now show that SM_1 is strictly stronger than WM_1 . Let I_A denote the indicator function of a set A ; i.e., $I_A(t) = 1$ if $t \in A$ and $I_A(t) = 0$ otherwise.

Example 12.3.1. *WM_1 convergence without SM_1 convergence.* To show that we can have $d_w(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ without $d_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, let $x \equiv (x^1, x^2) \in D([0, 2], \mathbb{R}^2)$ be defined by $x^1 = x^2 = 2I_{[1, 2]}$ and let $x_n^1 = 2I_{[1-n^{-1}, 2]}$ and $x_n^2 = I_{[1-n^{-1}, 1]} + 2I_{[1, 2]}$. The thin range of x is the set $\{(0, 0), (2, 2)\}$ plus the line segment $[(0, 0), (2, 2)]$ connecting those two

points, while the thin range of x_n is the set $\{(0, 0), (2, 1), (2, 2)\}$ plus the line segments $[(0, 0), (2, 1)]$ and $[(2, 1), (2, 2)]$. Since $(2, 1) \in \Gamma_{x_n}$ for all n but $(2, 1) \notin \Gamma_x$, we must have $d_s(x_n, x) \not\rightarrow 0$ as $n \rightarrow \infty$. On the other hand, the thick ranges of x and x_n , $n \geq 1$ all are $[0, 2] \times [0, 2]$. To demonstrate that $d_w(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, we construct suitable parametric representations. Let

$$\begin{aligned} r(0) &= 0, \quad r(1/3) = 1 = r(2/3), \quad r(1) = 2 \\ r_n(0) &= 0, \quad r_n(1/3) = 1 - n^{-1} = r_n((1 - n^{-1})/2), \\ r_n((1 + n^{-1})/2) &= 1 = r_n(2/3), \quad r_n(1) = 2 \\ u^1(0) &= 0 = u^1(1/3), \quad u^1(1/2) = 2 = u^1(1) \\ u_n^1(0) &= 0 = u_n^1(1/3), \quad u_n^1((1 - n^{-1})/2) = 2 = u_n^1(1) \\ u^2(0) &= 0 = u^2(1/3), \quad u^2(1/2) = 1, \quad u^2(2/3) = 2 = u^2(1) \\ u_n^2(0) &= 0 = u_n^2(1/3), \quad u_n^2((1 - n^{-1})/2) = 1 = u_n^2((1 + n^{-1})/2), \\ u_n^2(2/3) &= 2 = u_n^2(1) \end{aligned}$$

with $r, r_n, u^1, u_n^1, u^2, u_n^2$ defined by linear interpolation in the gaps. With this construction, $(u_n, r_n) \in \Pi_w(x_n)$ and $(u, r) \in \Pi_w(x)$, $\|r_n - r\| = n^{-1}$, $\|u_n^1 - u^1\| = 6n^{-1}$ and $\|u_n^2 - u^2\| = 3n^{-1}$. Hence,

$$d_w(x_n, x) \leq \|u_n - u\| \vee \|r_n - r\| = 6n^{-1} \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad \blacksquare$$

Example 12.3.2. d_w is not a metric. We now show that d_w in (3.8) is not a metric. For this purpose, we use a minor modification of Example 12.3.1. Let $x^1 = x^2 = 2I_{[1,2]}$ as before. For even n , let $x_n^1 = 2I_{[1-n^{-1},2]}$ and $x_n^2 = I_{[1-n^{-1},1]} + 2I_{[1,2]}$ as before. Then let $x_{2n+1}^1 = x_{2n}^2$ and $x_{2n+1}^2 = x_{2n}^1$. We show that $d_w(x_{2n}, x_{2n+1}) \not\rightarrow 0$ as $n \rightarrow \infty$ even though $d_w(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, contradicting the triangle inequality property of a metric. The thick range of x_n is $([0, 2] \times [0, 1]) \cup (\{2\} \times [1, 2])$ for n even and $([0, 1] \times [0, 2]) \cup ([1, 2] \times \{2\})$ for n odd. The points $(2, 1)$ and $(1, 2)$ appear for n even and odd, respectively, but are distance 1 from the other thick range. Any parametric representation must pass through $(2, 1, 1 - n^{-1})$ in $\mathbb{R}^2 \times [0, 2]$ for n even and $(1, 2, 1 - n^{-1})$ for n odd. However, for n odd (n even) all points on G_{x_n} are at least a distance 1 from $(2, 1, 1 - n^{-1})$ ($(1, 2, 1 - n^{-1})$). This example shows that we cannot find a constant K such that $d_w(x_1, x_2) \leq K d_p(x_1, x_2)$ for all $x_1, x_2 \in D$. $\blacksquare$

We now relate the metrics $d_{M_1} \equiv d_s$ and d_{J_1} for d_{J_1} in (3.2) of Section 3.3.

Theorem 12.3.2. (comparison of J_1 and M_1 metrics) *For each $x_1, x_2 \in D$,*

$$d_s(x_1, x_2) \leq d_{J_1}(x_1, x_2) .$$

Remark 12.3.2. *Uses of the M_1 topology.* The M_1 topology has not been used extensively. It was used by Whitt (1971b, 1980, 2000b), Wichura (1974), Avram and Taqqu (1989, 1992), Kella and Whitt (1990), Chen and Whitt (1993), Mandelbaum and Massey (1995), Harrison and Williams (1996), Puhalskii and Whitt (1997, 1998), Resnick and van der Berg (2000), O'Brien (2000) and no doubt a few others. ■

12.3.3. Properties of Parametric Representations

We conclude this section by further discussing strong parametric representations. We first indicate how to construct a parametric representation (u, r) of Γ_x for any $x \in D$.

Remark 12.3.3. *How to construct a parametric representation.* Let t_j , $j \geq 1$, be a list of the discontinuity points of x (of which there are finitely or countably infinite many). For each j , select a subinterval $[a_j, b_j] \subseteq [0, 1]$ and let $r(s) = t_j$ for $a_j \leq s \leq b_j$, $u(a_j) = x(t_j-)$, $u(b_j) = x(t_j)$ and $u(\alpha a_j + (1-\alpha)b_j) = \alpha u(a_j) + (1-\alpha)u(b_j)$, $0 < \alpha < 1$. For successive discontinuities, do this in an order-preserving way; i.e., if $t_i < t_j < t_k$, then we require that $b_i < a_j < b_j < a_k$. Let this be done for all j . Next, suppose that t is not a discontinuity point but is the limit of discontinuity points. If $t_j \downarrow t$ as $j \rightarrow \infty$ where $t_j \in \text{Disc}(x)$, then let $r(a) = t$ and $u(a) = \lim_{j \rightarrow \infty} x(t_j-)$, where $a = \lim_{j \rightarrow \infty} a_j$ with $r(a_j) = t_j$. Similarly, if $t_j \uparrow t$ as $j \rightarrow \infty$ where $t_j \in \text{Disc}(x)$, then let $r(b) = t$ and $u(b) = \lim_{j \rightarrow \infty} x(t_j)$, where $b = \lim_{j \rightarrow \infty} b_j$ with $r(b_j) = t_j$. Finally, there may remain open intervals (a, b) over which (u, r) is undefined. Since (u, r) is already defined at the endpoints a and b , let $r(\alpha a + (1-\alpha)b) = \alpha r(a) + (1-\alpha)r(b)$ and $u(\alpha a + (1-\alpha)b) = x(r(\alpha a + (1-\alpha)b))$ for $0 < \alpha < 1$. This construction makes (u, r) a one-to-one function. This construction also makes r a generalization of piecewise linear; i.e., there are finite or countably many subintervals $[a_j, b_j]$ over which r is constant and there are finite or countably many intervals (b_k, a_k) over which r is linear. The union of all those points (where r is constant or linear) is dense in $[0, 1]$. The function r is extended to all other points by continuity. ■

Remark 12.3.4. *Parametric representations need not be one-to-one.* We do not require that a parametric representation be a one-to-one function. For example, even if x is continuous at t , we could have $r(s) = t$ for $a \leq s \leq b$.

Then, necessarily, $u(s) = x(t)$, $a \leq s \leq b$. However, we get the same metric if the parametric representations (u, r) are required to be one-to-one with r nondecreasing, e.g., as done by Wichura (1974); see Remark 12.5.2 in Section 5. Skorohod (1956) only originally required that r be nondecreasing instead of (u, r) , without the one-to-one property, in his Definitions 2.2.4 and 2.2.5. However, from his remarks after 2.2.5, it is evident that he meant to require that (u, r) be nondecreasing as we have defined it. As stated, Skorohod's version of the M_1 topology with only r nondecreasing is actually the M_2 topology. ■

Example 12.3.3. *Need for monotonicity.* To see the importance of requiring that the parametric representation be nondecreasing, using the order on the graphs, let $x = I_{[1,2]}$, $x_n(1) = x_n(1 - 2n^{-1}) = x_n(2) = 1$ and $x_n(0) = x_n(1 - 3n^{-1}) = x_n(1 - n^{-1}) = 0$, with x_n defined by linear interpolation elsewhere. For these functions, $x_n \rightarrow x$ as $n \rightarrow \infty$ in the M_2 topology but not in the M_1 topology. If we did not require that parametric representations of x be nondecreasing in our M_1 definitions, then we would have $x_n \rightarrow x$ as $n \rightarrow \infty$ in the M_1 topology. To see this, we exhibit parametric representations. Let $u_n = u$, $n \geq 4$, and let

$$\begin{aligned} r(0) &= 0, \quad r(1/5) = r(4/5) = 1, \quad r(1) = 2 \\ u(0) &= u(1/5) = u(3/5) = 0, \quad u(2/5) = u(4/5) = u(2) = 1 \\ r_n(0) &= 0, \quad r_n(1/5) = 1 - 3n^{-1}, \quad r_n(2/5) = 1 - 2n^{-1}, \\ r_n(3/5) &= 1 - n^{-1}, \quad r_n(4/5) = 1, \quad r_n(1) = 2 \end{aligned}$$

with r, u, r_n and u_n defined by linear interpolation elsewhere. It is easy to see that $(u_n, r_n) \in \Pi_s(x_n)$, and $\|(u_n, r_n) - (u, r)\| = \|r_n - r\| = 3n^{-1}$, but $(u, r) \notin \Pi_s(x)$ because (u, r) fails to be nondecreasing, since it backtracks on the graph at $t = 1$. If r were only required to be nondecreasing, then we would have $(u, r) \in \Pi_s(x)$. ■

We now continue characterizing parametric representations. For $x \in D$, $t \in \text{Disc}(x)$ and $(u, r) \in \Pi_s(x)$, there exists a unique pair of points $s_l \equiv s_l(t, x)$ and $s_r \equiv s_r(t, x)$ such that $s_l < s_r$ and $r^{-1}(\{t\}) = [s_l, s_r]$, i.e.,

$$\begin{aligned} \text{(i)} \quad & r(s) < t \text{ for } s < s_l & (3.12) \\ \text{(ii)} \quad & r(s) = t \text{ for } s_l \leq s \leq s_r \\ \text{(iii)} \quad & r(s) > t \text{ for } s > s_r. \end{aligned}$$

We will exploit the fact that a parametric representation (u, r) in $\Pi_s(x)$ is *jump consistent*: for each $t \in \text{Disc}(x)$ and pair $s_l \equiv s_l(t, x) < s_r \equiv$

$s_r(t, x)$ such that (3.12) holds, there is a continuous nondecreasing function β_t mapping $[0, 1]$ onto $[0, 1]$ such that

$$u(s) = \beta_t \left(\frac{s - s_l}{s_r - s_l} \right) u(s_r) + \left[1 - \beta_t \left(\frac{s - s_l}{s_r - s_l} \right) \right] u(s_l) \quad \text{for } s_l \leq s \leq s_r. \quad (3.13)$$

Condition (3.13) means that u is defined within jumps by interpolation from the definition at the endpoints s_l and s_r , consistently over all coordinates. In particular, suppose that $t \in \text{Disc}(x^i)$. (Since $t \in \text{Disc}(x)$, we must have $t \in \text{Disc}(x^i)$ for some coordinate i .) Suppose that $x^i(t-) < x^i(t)$. Then we can let

$$\beta_t(s) = \frac{u^i(s) - u^i(s_l)}{u^i(s_r) - u^i(s_l)}. \quad (3.14)$$

We see that (3.13) and (3.14) are consistent in that

$$u^i(s) = \beta_t \left(\frac{s - s_l}{s_r - s_l} \right) u^i(s_r) + \left[1 - \beta_t \left(\frac{s - s_l}{s_r - s_l} \right) \right] u^i(s_l) \quad (3.15)$$

for β_t in (3.14). For another coordinate j , (3.13) and (3.14) imply that

$$u^j(s) = \left(\frac{u^i(s) - u^i(s_l)}{u^i(s_r) - u^i(s_l)} \right) u^j(s_r) + \left(\frac{u^i(s_r) - u^i(s)}{u^i(s_r) - u^i(s_l)} \right) u^j(s_l). \quad (3.16)$$

It is possible that $t \notin \text{Disc}(x^j)$, in which case $u^j(s) = u^j(s_l) = u^j(s_r)$ for all s , $s_l \leq s \leq s_r$.

We can further characterize the behavior of a strong parametric representation at a discontinuity point. For $x \in D$, $t \in \text{Disc}(x)$ and $(u, r) \in \Pi_s(x)$, there exists a unique set of four points $s_l \equiv s_l(t, x) \leq s'_l \equiv s'_l(t, x) < s'_r \equiv s'_r(t, x) \leq s_r \equiv s_r(t, x)$ such that (3.12) holds and

$$\begin{aligned} & \text{(i) } u(s) = u(s_l) \text{ for } s_l \leq s \leq s'_l, \\ & \text{(ii) for each } i, \text{ either } u^i(s_l) < u^i(s) < u^i(s_r), \\ & \quad \text{or } u^i(s_l) > u^i(s) > u^i(s_r) \text{ for } s'_l < s < s'_r, \\ & \text{(iii) } u(s) = u(s_r) \text{ for } s'_r \leq s \leq s_r. \end{aligned} \quad (3.17)$$

Let D_1 be the subset of D containing functions all of whose jumps occur in only one coordinate, i.e., the set of x such that, for each $t \in \text{Disc}(x)$ there exists one and only one $i \equiv i(t)$ such that $t \in \text{Disc}(x^i)$. (The coordinate i may depend on t .)

Lemma 12.3.3. (strong and weak parametric representations coincide on D_1) For each $x \in D_1$, $\Pi_s(x) = \Pi_w(x)$.

We now show that parametric representations are preserved under linear functions of the coordinates when $x \in \Pi_s(x)$. That is *not* true in $\Pi_w(x)$.

Lemma 12.3.4. (linear functions of parametric representations) *If $(u, r) \in \Pi_s(x)$, then $(\eta u, r) \in \Pi_s(\eta x)$ for any $\eta \in \mathbb{R}^k$.*

12.4. Local Uniform Convergence at Continuity Points

In this section we provide alternative characterizations of local uniform convergence at continuity points of a limit function. The non-uniform Skorohod topologies on D all imply local uniform convergence at continuity points of a limit function. They differ by their behavior at discontinuity points.

We first observe that pointwise convergence is weaker than local uniform convergence.

Example 12.4.1. *Pointwise convergence is weaker than local uniform convergence.* To see that pointwise convergence in D at all continuity points of the limit is strictly weaker than local uniform convergence at continuity points of the limit, let $x(t) = 0$, $0 \leq t \leq 2$, and $x_n = I_{[1+n^{-1}, 1+2n^{-1}]}$, $n \geq 1$. Then $x_n(t) \rightarrow x(t) = 0$ as $n \rightarrow \infty$ for all t , but $x_n(1 + n^{-1}) = 1 \not\rightarrow 0$ as $n \rightarrow \infty$, so we do not have local uniform convergence at $t = 1$. We also do not have $x_n \rightarrow x$ as $n \rightarrow \infty$ in D in any of the Skorohod topologies. ■

We start by defining two basic *uniform-distance functions*.

For $x_1, x_2 \in D$, $t \in [0, T]$ and $\delta > 0$, let

$$u(x_1, x_2, t, \delta) \equiv \sup_{0 \vee (t-\delta) \leq t_1 \leq (t+\delta) \wedge T} \{\|x_1(t_1) - x_2(t_1)\|\} , \quad (4.1)$$

$$v(x_1, x_2, t, \delta) \equiv \sup_{0 \vee (t-\delta) \leq t_1, t_2 \leq (t+\delta) \wedge T} \{\|x_1(t_1) - x_2(t_2)\|\} , \quad (4.2)$$

We also define an *oscillation function*. For $x \in D$, $t \in [0, T]$ and $\delta > 0$, let

$$\bar{v}(x, t, \delta) \equiv \sup_{0 \vee (t-\delta) \leq t_1 \leq t_2 \leq (t+\delta) \wedge T} \{\|x(t_1) - x(t_2)\|\} . \quad (4.3)$$

We next define oscillation functions that we will use with the M_1 topologies. They use the distance $\|z - A\|$ between a point z and a subset A in $\mathbb{R}^k$ defined in (5.3) in Section 11.5. The SM_1 and WM_1 topologies use the

standard and product segments in (3.1) and (3.2). For each $x \in D$, $t \in [0, T]$ and $\delta > 0$, let

$$w_s(x, t, \delta) \equiv \sup_{0 \vee (t-\delta) \leq t_1 < t_2 < t_3 \leq (t+\delta) \wedge T} \{\|x(t_2) - [x(t_1), x(t_3)]\|\} \quad (4.4)$$

and

$$w_w(x, t, \delta) \equiv \sup_{0 \vee (t-\delta) \leq t_1 < t_2 < t_3 \leq (t+\delta) \wedge T} \{\|x(t_2) - [[x(t_1), x(t_3)]]\|\} \quad (4.5)$$

We now turn to the M_2 topology, which we will be studying in Sections 12.10 and 12.11. We define two uniform-distance functions. We use $\bar{w}$ as opposed to w to denote an M_2 uniform-distance function. Just as with the M_1 topologies, the SM_2 and WM_2 topologies use the standard and product segments in (3.1) and (3.2). For $x_1, x_2 \in D$, let

$$\bar{w}_s(x_1, x_2, t, \delta) \equiv \sup_{0 \vee (t-\delta) \leq t_1 \leq (t+\delta) \wedge T} \{\|x_1(t_1) - [x_2(t-), x_2(t)]\|\} \quad (4.6)$$

$$\bar{w}_w(x_1, x_2, t, \delta) \equiv \sup_{0 \vee (t-\delta) \leq t_1 \leq (t+\delta) \wedge T} \{\|x_1(t_1) - [[x_2(t-), x_2(t)]]\|\} \quad (4.7)$$

It is easy to establish the following relations among the uniform-distance and oscillation functions.

Lemma 12.4.1. (inequalities for uniform-distance and oscillation functions)
For all $x, x_n \in D$, $t \in [0, T]$ and $\delta > 0$,

$$u(x_n, x, t, \delta) \leq v(x_n, x, t, \delta) \leq u(x_n, x, t, \delta) + \bar{v}(x, t, \delta) ,$$

$$w_w(x_n, t, \delta) \leq w_s(x_n, t, \delta) \leq \bar{v}(x_n, t, \delta) \leq 2v(x_n, x, t, \delta) + \bar{v}(x, t, \delta) ,$$

$$\bar{w}_w(x_n, x, t, \delta) \leq \bar{w}_s(x_n, x, t, \delta) \leq v(x_n, x, t, \delta) \leq 2\bar{w}_w(x_n, x, t, \delta) + \bar{v}(x, t, \delta) .$$

Since the M_1 -oscillation functions $w_s(x_n, t, \delta)$ and $w_w(x_n, t, \delta)$ do not contain the limit x , their convergence to 0 as $n \rightarrow \infty$ and then $\delta \downarrow 0$ does not directly imply local uniform convergence at a continuity point of a prospective limit function x .

Example 12.4.2. *Characterizations of local uniform convergence at continuity points.* We show that it is possible to have $t \notin \text{Disc}(x)$, $x_n(t) \rightarrow x(t)$ as $n \rightarrow \infty$ and

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w_s(x_n, t, \delta) = 0$$

without having

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} v(x_n, x, t, \delta) = 0.$$

That occurs for $t = 1$ when $x(t) = 0$, $0 \leq t \leq 2$, and $x_n = I_{[1+n^{-1}, 2]}$, $n \geq 1$. In this example, we have $\bar{v}(x, t, \delta) = 0$ and $w_s(x_n, t, \delta) = 0$ for all n , t and $\delta > 0$, but $v(x_n, x, 1, \delta) = 1$ for $n > 1/\delta$. ■

We relate convergence of $w_s(x_n, t, \delta)$ and $w_w(x_n, t, \delta)$ to 0 as $n \rightarrow \infty$ and $\delta \downarrow 0$ to local uniform convergence by requiring pointwise convergence in a neighborhood of t ; see (vi) in Theorem 12.4.1 below.

Theorem 12.4.1. (characterizations of local uniform convergence at continuity points) *If $t \notin \text{Disc}(x)$, then the following are equivalent:*

$$(i) \quad \lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} u(x_n, x, t, \delta) = 0, \quad (4.8)$$

$$(ii) \quad \lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} v(x_n, x, t, \delta) = 0, \quad (4.9)$$

$$(iii) \quad \lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{w}_s(x_n, x, t, \delta) = 0, \quad (4.10)$$

$$(iv) \quad \lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{w}_w(x_n, x, t, \delta) = 0, \quad (4.11)$$

(v) $x_n(t_1) \rightarrow x(t_1)$ for all t_1 in a dense subset of a neighborhood of t (including 0 if $t = 0$ or T if $t = T$) and

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w_s(x_n, t, \delta) = 0,$$

(vi) $x_n(t_1) \rightarrow x(t_1)$ for all t_1 in a dense subset of a neighborhood of t (including 0 if $t = 0$ or T if $t = T$) and

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w_w(x_n, t, \delta) = 0. \quad (4.12)$$

We now show that local uniform convergence at all points in a compact interval implies uniform convergence over the compact interval.

Lemma 12.4.2. (local uniform convergence everywhere in a compact interval) *If (4.8) holds for all $t \in [a, b]$, then*

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \sup_{0 \vee (a-\delta) \leq t \leq (b+\delta) \wedge T} \{\|x_n(t) - x(t)\|\} = 0.$$

12.5. Alternative Characterizations of M_1 Convergence

We now give alternative characterizations of SM_1 and WM_1 convergence.

12.5.1. SM_1 Convergence

We first give several alternative characterizations of SM_1 -convergence (or, equivalently, d_s -convergence) in D , one being a minor variant of the original one involving an oscillation function established by Skorohod (1956). Another one – (v) below – involves only the local behavior of the functions. It helps us establish sufficient conditions to have $d_s((x_n, y_n), (x, y)) \rightarrow 0$ in $D([0, T], \mathbb{R}^{k+l})$ when $d_s(x_n, x) \rightarrow 0$ in $D([0, T], \mathbb{R}^k)$ and $d_s(y_n, y) \rightarrow 0$ in $D([0, T], \mathbb{R}^l)$; see Section 12.6. For the SM_1 topology, we define another oscillation function. For any $x_1, x_2 \in D$ and $\delta > 0$, let

$$w_s(x, \delta) \equiv \sup_{0 \leq t \leq T} w_s(x, t, \delta) , \quad (5.1)$$

for $w_s(x, t, \delta)$ in (4.4). We include the proof here, except for the supporting lemmas, which are proved in the Internet Supplement.

Theorem 12.5.1. (characterizations of SM_1 convergence) *The following are equivalent characterizations of convergence $x_n \rightarrow x$ as $n \rightarrow \infty$ in (D, SM_1) :*

- (i) *For any $(u, r) \in \Pi_s(x)$, there exists $(u_n, r_n) \in \Pi_s(x_n)$, $n \geq 1$, such that*

$$\|u_n - u\| \vee \|r_n - r\| \rightarrow 0 \quad \text{as } n \rightarrow \infty . \quad (5.2)$$

- (ii) *There exist $(u, r) \in \Pi_s(x)$ and $(u_n, r_n) \in \Pi_s(x_n)$ for $n \geq 1$ such that (5.2) holds.*

- (iii) *$d_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$; i.e., for all $\epsilon > 0$ and all sufficiently large n , there exist $(u, r) \in \Pi_s(x)$ and $(u_n, r_n) \in \Pi_s(x_n)$ such that*

$$\|u_n - u\| \vee \|r_n - r\| < \epsilon .$$

- (iv) *$x_n(t) \rightarrow x(t)$ as $n \rightarrow \infty$ for each t in a dense subset of $[0, T]$ including 0 and T , and*

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w_s(x_n, \delta) = 0 \quad (5.3)$$

for $w_s(x, \delta)$ in (5.1) and $w_s(x, t, \delta)$ in (4.4).

(v) $x_n(T) \rightarrow x(T)$ as $n \rightarrow \infty$; for each $t \notin \text{Disc}(x)$,

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} v(x_n, x, t, \delta) = 0 \quad (5.4)$$

for $v(x_1, x_2, t, \delta)$ in (4.2); and, for each $t \in \text{Disc}(x)$,

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w_s(x_n, t, \delta) = 0 \quad (5.5)$$

for $w_s(x, t, \delta)$ in (4.4).

(vi) For all $\epsilon > 0$, , there exist integers m and n_1 , a finite ordered subset A of Γ_x of cardinality m as in (3.9) and, for all $n \geq n_1$, finite ordered subsets A_n of Γ_{x_n} of cardinality m such that, for all $n \geq n_1$, $\hat{d}(A, \Gamma_x) < \epsilon$, $\hat{d}(A_n, \Gamma_{x_n}) < \epsilon$ for $\hat{d}$ in (3.10) and $d^*(A, A_n) < \epsilon$, where

$$d^*(A, A_n) \equiv \max_{1 \leq i \leq m} \{ \| (z_i, t_i) - (z_{n,i}, t_{n,i}) \| : (z_i, t_i) \in A, (z_{n,i}, t_{n,i}) \in A_n \}. \quad (5.6)$$

In preparation for the proof of Theorem 12.5.1, we establish some preliminary results. We first show that SM_1 convergence implies local uniform convergence at all continuity points.

Lemma 12.5.1. (local uniform convergence) *If $d_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, then (4.9) holds for each $t \notin \text{Disc}(x)$.*

We next relate the modulus w_s applied to x and the modulus applied to corresponding points on the graph Γ_x . The following lemma is established in the proof of Skorohod's (1956) 2.4.1.

Lemma 12.5.2. (extending the modulus from a function to its graph) *If $(z_1, t_1), (z_2, t_2), (z_3, t_3) \in \Gamma_x$ with $0 \vee (t - \delta) \leq t_1 < t_2 < t_3 \leq (t + \delta) \wedge T$, then $\|z_2 - [z_1, z_3]\| \leq w_s(x, \delta)$.*

Lemma 12.5.3. (asymptotic negligibility of the modulus) *For any $x \in D$, $w_s(x, \delta) \downarrow 0$ as $\delta \downarrow 0$.*

Proof of Theorem 12.5.1. The implications (i)→(ii)→(iii) are trivial. We establish the others exploiting transitivity.

(iii)→(iv). First, the convergence $x_n(T) \rightarrow x(T)$ is assumed directly. Next, by Lemma 12.5.1, if $d_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, then $x_n(t) \rightarrow x(t)$ for all $t \in \text{Disc}(x)^c$, which is a dense subset of $[0, T]$. We now want to show that, for any $\epsilon > 0$, there exists n_0 and δ such that $w_s(x_n, \delta) < \epsilon$ for all $n \geq n_0$. For $x \in D$ and $\epsilon > 0$ given, start by choosing η so that $w_s(x, \eta) < \epsilon/2$, which we can do by Lemma 12.5.3. Then apply (iii) to choose n_0 so that $(u_n, r_n) \in \Pi_s(x_n)$, $(u, r) \in \Pi_s(x)$ and

$$\|u_n - u\| \vee \|r_n - r\| < (\epsilon \wedge \eta)/4 \quad \text{for } n \geq n_0 .$$

Suppose that $(t - \delta) \vee 0 \leq t_1 < t_2 \leq t_3 < (t + \delta) \wedge T$. Let $s_{n,i}$ be such that $r_n(s_{n,i}) = t_i$ and $u_n(s_{n,i}) = x_n(t_i)$ for $i = 1, 2, 3$ and all n . Then, apply Lemma 12.5.2 to obtain, for $n \geq n_0$,

$$\begin{aligned} \|x_n(t_2) - [x_n(t_1), x_n(t_3)]\| &= \|u_n(s_{n,2}) - [u_n(s_{n,1}), u_n(s_{n,3})]\| \\ &\leq \|u(s_{n,2}) - [u(s_{n,1}), u(s_{n,3})]\| + 2\|u_n - u\| \\ &\leq w_s(x, \delta + 2(\eta \wedge \epsilon)/4) + 2((\eta \wedge \epsilon)/4) \\ &\leq w_s(x, \delta + (\eta/2)) + \epsilon/2 , \end{aligned}$$

so that, for $\delta < \eta/2$ and $n \geq n_0$, $w_s(x_n, \delta) < \epsilon$.

(iv)→(vi). First, for $\epsilon > 0$ given, apply (iv) to find $\eta < \epsilon/16$ and n_0 such that $w_s(x_n, \eta) < \epsilon/32$ for $n \geq n_0$. Next find a finite set A of points (z_i, t_i) in Γ_x with

$$(x(0), 0) = (z_1, t_1) < (z_2, t_2) < \cdots < (z_m, t_m) = (x(T), T) ,$$

using the order defined on Γ_x below (3.3), where for each i , either $t_i \in \text{Disc}(x, \epsilon/2)$ or $t_i \in S$, with $\text{Disc}(x, \epsilon/2)$ being as in (2.4) and S being a subset of $[0, T]$ including 0, T and the points in $\text{Disc}(x, \epsilon/2)^c$ at which x_n converges pointwise to x . Use the left and right limits of x to include in A for each $t \in \text{Disc}(x, \epsilon/2)$ points $t' \equiv t'(t)$ and $t'' = t''(t)$ in S such that $t' < t < t''$, t' is greater than all elements of $\text{Disc}(x, \epsilon/2)$ less than t , t'' is less than all elements of $\text{Disc}(x, \epsilon/2)$ greater than t , $|t' - t| < \eta$, $|t'' - t| < \eta$, $\|x(t') - x(t-)\| < \epsilon/32$ and $\|x(t'') - x(t)\| < \epsilon/32$. In addition, assume that $|t_{i+1} - t_i| < \eta$ for all i and $\hat{d}(A, \Gamma_x) < \epsilon/2$, for which we apply Lemma 12.3.1. Moreover, if $t \in \text{Disc}(x, \epsilon/2)$ and

$$t_r < t_{r+1} = t = t_{r+2} = \cdots = t_{r+j} < t_{r+j+1} , \quad (5.7)$$

then we require that $\|z_{r+1} - x(t-)\| > \epsilon/4$, $\|z_{r+j} - x(t)\| > \epsilon/4$ and $\|z_{r+i+1} - z_{r+i}\| > \epsilon/4$ for all i , $1 \leq i \leq j-1$. Since $\hat{d}(A, \Gamma_x) < \epsilon/2$, we also have the upper bound $\|z_{r+i+1} - z_{r+i}\| < \epsilon/2$. For $t_i \in S \cap A$, let $z_i = x(t_i)$. Now, for all $t_i \in S \cap A$, let $n_1 \geq n_0$ be such that $\|x_n(t_i) - x(t_i)\| < \epsilon/32$ for all i and $n \geq n_1$, using (iv). We now want to construct the subset A_n of Γ_{x_n} . First for all $t_i \in S \cap A$, let $(z_{n,i}, t_{n,i}) = (x_n(t_i), t_i)$. Now we consider time points in $\text{Disc}(x, \epsilon/2)$. By the construction above, given (5.7),

$$\|[x(t_r), x(t_{r+j+1})] - [x_n(t_r), x_n(t_{r+j+1})]\| < \epsilon/32$$

and

$$\|[x(t_r), x(t_{r+j+1})] - [x(t-), x(t)]\| < \epsilon/32. \quad (5.8)$$

Since $w_s(x_n, \eta) < \epsilon/32$, for each (r, i) there is a point $(z_{n,r+i}, t_{n,r+i}) \in \Gamma_{x_n}$ such that

$$\|z_{n,r+i} - z_{r+i}\| < 3\epsilon/32 \quad \text{and} \quad |t_{n,r+i} - t| < \eta < \epsilon/16. \quad (5.9)$$

Moreover, we must have $(z_{n,r+i+1}, t_{n,r+i+1}) > (z_{n,r+i}, t_{n,r+i})$ for $0 \leq i \leq j$. For $i = 0$ and $i = j$, we can conclude that $t_r < t < t_{r+j+1}$. For other i , a reversal of order can occur only if $w_s(x_n, t, \eta) > \epsilon/16$ because the construction implies that $\|z_{n,r+i+1} - z_{n,r+i}\| > \epsilon/16$, but that is prohibited by the condition that $w_s(x_n, t, \eta) < \epsilon/32$. Hence, the set of points A_n is ordered properly. Moreover, the construction yields $d^*(A, A_n) < \epsilon/16$. Finally, it remains to bound $\hat{d}(A_n, \Gamma_{x_n})$ for $n \geq n_1$. Consider (z_n, t_n) such that $(z_{n,i}, t_{n,i}) < (z_n, t_n) < (z_{n,i+1}, t_{n,i+1})$. Since $\|z_{n,i} - z_i\| < 3\epsilon/32$ for all i and $\|z_{i+1} - z_i\| < \hat{d}(A, \Gamma_x) < \epsilon/2$, $\|z_{n,i+1} - z_{n,i}\| < 5\epsilon/8$ by the triangle inequality. Since $w_s(x_n, \eta) < \epsilon/32$, invoking Lemma 12.5.2, we have

$$\|(z_n, t_n) - [(z_{n,i}, t_{n,i}), (z_{n,i+1}, t_{n,i+1})]\| < \epsilon/32,$$

so that

$$\|z_n - z_{n,i}\| \vee \|z_n - z_{n,i+1}\| < 21\epsilon/32 < \epsilon$$

and $|t_{n,i} - t_{n,i+1}| < 2\eta < \epsilon/8$. Hence $\hat{d}(A_n, \Gamma_{x_n}) < \epsilon$ for $n \geq n_1$, so that the proof is complete.

(vi)→(i). Suppose that the conditions in (vi) hold and $\epsilon > 0$ is given. Let $(u, r) \in \Pi_s(x)$ and $(u_n, r_n) \in \Pi_s(x_n)$, $n \geq 1$, be arbitrary parametric representations. Let $s_1 = 0 < s_2 < \dots < s_m = 1$ and $s_{n,1} = 0 < s_{n,2} < \dots < s_{n,m} = 1$ be points such that $(u(s_i), r(s_i)) = (z_i, t_i) \in A$ and $(u_n(s_{n,i}), r_n(s_{n,i})) = (z_{n,i}, t_{n,i}) \in A_n$ for $1 \leq i \leq m$. Let $\lambda_n : [0, 1] \rightarrow [0, 1]$ be a continuous nondecreasing function such that $\lambda_n(s_i) = s_{n,i}$ for each i

and n . We will show that $(u_n \circ \lambda_n, r_n \circ \lambda_n)$ is a parametric representation of Γ_{x_n} for each n such that

$$\|u_n \circ \lambda_n - u\| \vee \|r_n \circ \lambda_n - r\| < 3\epsilon \quad \text{for } n \geq n_1. \quad (5.10)$$

Property (5.10) holds because, for $s_i \leq s \leq s_{i+1}$, $\lambda_n(s_i) = s_{n,i} \leq \lambda_n(s) \leq s_{n,i+1} = \lambda_n(s_{i+1})$ and

$$\begin{aligned} & \|u_n \circ \lambda_n(s) - u(s)\| \vee \|r_n \circ \lambda_n(s) - r(s)\| \\ & \leq \| (u_n \circ \lambda_n(s), r_n \circ \lambda_n(s)) - (u_n(s_{n,i}), r_n(s_{n,i})) \| \vee \| (u_n \circ \lambda_n(s), r_n \circ \lambda_n(s)) \\ & \quad - (u_n(s_{n,i+1}), r_n(s_{n,i+1})) \| \\ & + \| (u(s), r(s)) - (u(s_i), r(s_i)) \| \vee \| (u(s), r(s)) - (u(s_{i+1}), r(s_{i+1})) \| \\ & + \| (u_n(s_{n,i}), r_n(s_{n,i})) - (u(s_i), r(s_i)) \| \\ & \leq \hat{d}(A_n, \Gamma_{x_n}) + \hat{d}(A, \Gamma_x) + d^*(A, A_n) \leq 3\epsilon. \end{aligned}$$

(v) $\rightarrow$ (iv). First, the convergence $x_n(T) \rightarrow x(T)$ is assumed directly. Next, (5.4) implies that $x_n(t) \rightarrow x(t)$ as $n \rightarrow \infty$ for each $t \notin \text{Disc}(x)$. Since $[0, T] - \text{Disc}(x)$ is a dense subset of $[0, T]$, the first part of (iv) is established. Condition (5.4) also implies that (5.5) holds for each $t \notin \text{Disc}(x)$ by Theorem 12.4.1. Finally, we show that (5.5) for each $t \in [0, T]$ implies (5.3). Condition (5.5) for each t implies that for each $\epsilon > 0$ and t , there is $\delta \equiv \delta(t)$ and $n(t, \epsilon, \delta)$ such that $w_s(x_n, t, \delta) < \epsilon$ for all $n \geq n(t, \epsilon, \delta)$. Now suppose that (5.3) does not hold. Then there must exist $\epsilon > 0$ such that for all $\delta > 0$ there is a sequence $\{t_k\}$ of points in $[0, T]$ and a sequence of integers n_k such that $n_k \rightarrow \infty$ and $w_s(x_{n_k}, t_k, \delta/2) > \epsilon$ for all k . However, the sequence $\{t_k\}$ has a subsequence $\{t_{k_j}\}$ with $t_{k_j} \rightarrow t \in [0, T]$ as $k_j \rightarrow \infty$. Thus, for all k_j suitably large,

$$w_s(x_{n_{k_j}}, t, \delta) > w_s(x_{n_{k_j}}, t_{n_{k_j}}, \delta/2) > \epsilon,$$

which is a contradiction, so that (5.3) must in fact hold.

(iii) + (iv) $\rightarrow$ (v). By Lemma 12.5.1, (iii) implies (5.4) for each $t \in \text{Disc}(x)^c$. Trivially, (iv) implies (5.3), which in turn implies (5.5). ■

Remark 12.5.1. *Connection to Skorohod (1956).* Part (iv) of Theorem 12.5.1 is essentially Skorohod's (1956) original characterization, established in his 2.4.1. Instead of (5.1) with (4.4), Skorohod (1956) actually considered (5.1) with $w_s(x, t, \delta)$ replaced by

$$w'_s(x, t, \delta) \equiv \sup_{0 \vee (t-\delta) \leq t_1 \leq t \leq t_3 \leq (t+\delta) \wedge T} \{ \|x(t) - [x(t_1), x(t_3)]\| \}, \quad (5.11)$$

but when the supremum over $t \in [0, T]$ is applied, w_s and w'_s are equivalent. In particular, clearly $w'_s(x, t, \delta) \leq w_s(x, t, \delta)$ for each t . On the other hand, if $w_s(x, t, \delta) > \epsilon$ for all t , then $w'_s(x, t, 2\delta) > \epsilon$ for all t . Hence (iv) is equivalent to Skorohod's original characterization. We have introduced $w_s(x, t, \delta)$ in (4.4) in order to get characterization (v) in Theorem 12.5.1. We cannot use Skorohod's (5.11) instead of (4.4) in characterization (v) in Theorem 12.5.1, because it does not rule out multiple large oscillations on the same side of t . ■

Remark 12.5.2. *Possibility of using one-to-one parametric representations.* The proof of the implication (vi) $\rightarrow$ (i) shows that the SM_1 topology is unaltered if all the parametric representations are required to be one-to-one functions from $[0, 1]$ onto the graph. In the proof we would then let the transformations λ_n be homeomorphisms of $[0, 1]$, so that $(u_n \circ \lambda_n, r_n \circ \lambda_n)$ become one-to-one functions. ■

We can apply Theorem 12.5.1 to develop a simple criterion for M_1 convergence for monotone functions.

Corollary 12.5.1. (the case of monotone functions) *If x_n is monotone for each n , then $d_s(x_n, x) \rightarrow 0$ for $x \in D$ if and only if $x_n(t) \rightarrow x(t)$ for all t in a dense subset of $[0, T]$ including 0 and T .*

Proof. Apply Theorem 12.5.1 (iv). Note that condition (5.3) always holds for monotone functions. ■

12.5.2. WM_1 Convergence

We now establish an analog of Theorem 12.5.1 for the WM_1 topology. Several alternative characterizations of WM_1 convergence will follow directly from Theorem 12.5.1 because we will show that convergence $x_n \rightarrow x$ as $n \rightarrow \infty$ in WM_1 is equivalent to $d_p(x_n, x) \rightarrow 0$. To treat the WM_1 topology, we define another oscillation function. Let

$$w_w(x, \delta) \equiv \sup_{0 \leq t \leq T} w_w(x, t, \delta) \quad (5.12)$$

for $w_w(x, t, \delta)$ in (4.5). Recall that $w_w(x, t, \delta)$ in (4.5) is the same as $w_s(x, t, \delta)$ in (4.4) except it has the product segment $[[x(t_1), x(t_3)]]$ in (3.2) instead of the standard segment $[x(t_1), x(t_3)]$ in (3.1).

Paralleling Definition 12.3.1, let an ordered subset A of G_x of cardinality m be such that (3.9) holds, but now with the order being the order on G_x . Paralleling (3.10), let the *order-consistent distance* between A and G_x be

$$\hat{d}(A, G_x) \equiv \sup\{\|(z, t) - (z_i, t_i)\| \vee \|(z, t) - (z_{i+1}, t_{i+1})\| : (z, t) \in G_x\} \quad (5.13)$$

with the supremum being over all $(z, t) \in G_x$ such that $(z_i, t_i) \leq (z, t) \leq (z_{i+1}, t_{i+1})$ for all i , $1 \leq i \leq m - 1$.

Theorem 12.5.2. (characterizations of WM_1 convergence) *The following are equivalent characterizations of $x_n \rightarrow x$ as $n \rightarrow \infty$ in (D, WM_1) :*

- (i) $d_w(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$.
- (ii) $d_p(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$.
- (iii) $x_n(t) \rightarrow x(t)$ as $n \rightarrow \infty$ for each t in a dense subset of $[0, T]$ including 0 and T , and

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w_w(x_n, \delta) = 0. \quad (5.14)$$

- (iv) $x_n(T) \rightarrow x(T)$ as $n \rightarrow \infty$; for each $t \notin \text{Disc}(x)$,

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} v(x_n, x, t, \delta) = 0 \quad (5.15)$$

for $v(x_n, x, t, \delta)$ in (4.2); and, for each $t \in \text{Disc}(x)$,

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w_w(x_n, t, \delta) = 0 \quad (5.16)$$

for $w_w(x_n, t, \delta)$ in (4.5).

- (v) for all $\epsilon > 0$ and all n sufficiently large, there exist finite ordered subsets A of G_x (in general depending on n) and A_n of G_{x_n} of common cardinality such that $\hat{d}(A, G_x) < \epsilon$, $\hat{d}(A_n, G_{x_n}) < \epsilon$ and $d^*(A, A_n) < \epsilon$ for $\hat{d}$ in (5.13) and d^* in (5.6).

Example 12.5.1. *Need for changing parametric representations.* In general, there is no analog of characterizations (i) and (ii) in Theorem 12.5.1 for the parametric representations in $\Pi_w(x)$ and $\Pi_w(x_n)$; i.e., if $d_w(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, there need not exist $(u, r) \in \Pi_w(x)$ and $(u_n, r_n) \in \Pi_w(x_n)$ such that (5.2) holds. To see this, let $x^1 = x^2 = I_{[1,2]}$, $x_{2n+1}^1 = x_{2n}^2 = I_{[1-n^{-1},2]}$ and $x_{2n+1}^2 = x_{2n}^1 = I_{[1+n^{-1},2]}$ for $n \geq 2$. Property (i) of Theorem 12.5.2 holds, but different parametric representations of x are needed for even and odd n . ■

12.6. Strengthening the Mode of Convergence

In this section we apply the characterizations of M_1 convergence in Sections 12.3 and 12.5 to establish conditions under which the mode of convergence can be strengthened: We seek conditions under which WM_1 convergence can be replaced by SM_1 convergence. We use the following Lemma.

Lemma 12.6.1. (modulus bound for (x_n, y_n)) *For $x_n \in D([0, T], \mathbb{R}^k)$, $y_n, y \in D([0, T], \mathbb{R}^l)$, $t \in [0, T]$ and $\delta > 0$,*

$$w_s((x_n, y_n), t, \delta) \leq w_s(x_n, t, \delta) + 2v(y_n, y, t, \delta).$$

Theorem 12.6.1. (extending SM_1 convergence to product spaces) *Suppose that $d_s(x_n, x) \rightarrow 0$ in $D([0, T], \mathbb{R}^k)$ and $d_s(y_n, y) \rightarrow 0$ in $D([0, T], \mathbb{R}^l)$ as $n \rightarrow \infty$. If*

$$Disc(x) \cap Disc(y) = \phi.$$

then

$$d_s((x_n, y_n), (x, y)) \rightarrow 0 \text{ in } D([0, T], \mathbb{R}^{k+l}) \text{ as } n \rightarrow \infty.$$

Proof. We use characterization (v) in Theorem 12.5.1. First, for each $t \notin Disc((x, y))$, $t \notin Disc(x) \cup Disc(y)$, (5.4) holds and

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} v(y_n, y, \delta, t) = 0, \quad (6.1)$$

which implies that

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} v((x_n, y_n), (x, y), \delta, t) = 0.$$

Now, for each $t \in Disc(x)$, (5.5) and (6.1) hold (because $Disc(x) \cap Disc(y) = \phi$). Thus, for those t , by Lemma 12.6.1,

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w_s((x_n, y_n), t, \delta) = 0. \quad (6.2)$$

By the same reasoning (6.2) also holds for each $t \in Disc(y)$, so that (6.2) holds for all $t \in Disc((x, y)) = Disc(x) \cup Disc(y)$. ■

Remark 12.6.1. *The discontinuity condition is not necessary.* The discontinuity condition $Disc(x) \cap Disc(y) = \emptyset$ in Theorem 12.6.1 is not necessary. To see that, note that if $x_n \rightarrow x$ as $n \rightarrow \infty$ in $D([0, T], \mathbb{R}^k)$, then $(x_n, x_n) \rightarrow (x, x)$ as $n \rightarrow \infty$ in $D([0, T], \mathbb{R}^{2k})$. However, some condition is needed, as can be seen from the fact that the WM_1 topology is strictly weaker than the SM_1 topology on $D([0, T], \mathbb{R}^k)$ for $k > 1$, as shown by Example 12.3.1. ■

Remark 12.6.2. *The J_1 and M_2 analogs.* Analogs of Theorem 12.6.1 hold in the J_1 and M_2 topologies. For J_1 , see Propositions 2.1 (a) and 2.2 (b) on p. 301 of Jacod and Shiryaev (1987). For M_2 , see Theorem 12.11.3 below. ■

As in Lemma 12.3.3, let $D_1 \equiv D_1([0, T], \mathbb{R}^k)$ be the subset of x in D with discontinuities in only one coordinate at a time; i.e., $x \in D_1$ if $x^i(t-) \neq x^i(t)$ for at most one coordinate i for each t . (The coordinate $i \equiv i(t)$ may depend upon t .)

Corollary 12.6.1. (from WM_1 convergence to SM_1 convergence when the limit is in D_1) *If $d_p(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ and $x \in D_1$, then $d_s(x_n, x) \rightarrow 0$.*

Example 12.3.3 shows that it is not enough to have $x \in D_s$ in Corollary 12.6.1.

12.7. Characterizing Convergence with Mappings

The strong topology SM_1 differs from the weak topology WM_1 by the behavior of linear functions of the coordinates. Example 12.3.1 shows that linear functions of the coordinates are not continuous in the product topology (there $(x_n^1 - x_n^2) \not\rightarrow (x^1 - x^2)$ as $n \rightarrow \infty$), but they are in the strong topology, as we now show. Note that there is no subscript on d on the left in (7.1) below because ηx is real valued.

Theorem 12.7.1. (Lipschitz property of linear functions of the coordinate functions) *For any $x_1, x_2 \in D([0, T], \mathbb{R}^k)$ and $\eta \in \mathbb{R}^k$,*

$$d(\eta x_1, \eta x_2) \leq (\|\eta\| \vee 1) d_s(x_1, x_2) . \quad (7.1)$$

Example 12.7.1. *Difficulties with the weak topology.* To see that $(\eta z, t)$ need not be an element of $\Gamma_{\eta x}$ when $(z, t) \in G_x$ and that $(\eta u, r)$ need not be an element of $\Pi(\eta x)$ when $(u, r) \in \Pi_w(x)$, let $x^1 = x^2 = I_{[1, 2]}$ and consider $\eta x = x^1 - x^2$. The flexibility allowed by G_x allows $(z, t) \in G_x$ with $\eta z \neq 0$ and $(u, r) \in \Pi_w(x)$ with $u(s) \neq 0$. ■

We now obtain a sufficient condition for addition to be continuous on $(D, d_s) \times (D, d_s)$, which is analogous to the J_1 result in Theorem 4.1 of Whitt (1980).

Corollary 12.7.1. (SM_1 -continuity of addition) *If $d_s(x_n, x) \rightarrow 0$ and $d_s(y_n, y) \rightarrow 0$ in $D([0, T], \mathbb{R}^k)$ and*

$$Disc(x) \cap Disc(y) = \emptyset,$$

then

$$d_s(x_n + y_n, x + y) \rightarrow 0 \quad \text{in } D([0, T], \mathbb{R}^k).$$

Proof. First apply Theorem 12.6.1 to get $d_s((x_n, y_n), (x, y)) \rightarrow 0$ in $D([0, T], \mathbb{R}^{2k})$. Then apply Theorem 12.7.1. ■

Remark 12.7.1. *Measurability of addition.* The measurability of addition on $(D, d_s) \times (D, d_s)$ holds because the Borel σ -field coincides with the Kolmogorov σ -field. It also follows from part of the proof of Theorem 4.1 of Whitt (1980). ■

In Theorem 12.7.1 we showed that linear functions of the coordinates are Lipschitz in the SM_1 metric. We now apply Theorem 12.5.1 to show that convergence in the SM_1 topology is characterized by convergence of all such linear functions of the coordinates.

Theorem 12.7.2. (characterization of SM_1 convergence by convergence of all linear functions) *There is convergence $x_n \rightarrow x$ in $D([0, T], \mathbb{R}^k)$ as $n \rightarrow \infty$ in the SM_1 topology if and only if $\eta x_n \rightarrow \eta x$ in $D([0, T], \mathbb{R}^1)$ as $n \rightarrow \infty$ in the M_1 topology for all $\eta \in \mathbb{R}^k$.*

We can get convergence of sums under more general conditions than in Corollary 12.7.1. It suffices to have the jumps of x^i and y^i have common sign for all i . We can express this property by the condition

$$(x^i(t) - x^i(t-))(y^i(t) - y^i(t-)) \geq 0 \tag{7.2}$$

for all t , $0 \leq t \leq T$, and all i , $1 \leq i \leq k$.

Theorem 12.7.3. (continuity of addition at limits with jumps of common sign) *If $x_n \rightarrow x$ and $y_n \rightarrow y$ in $D([0, T], \mathbb{R}^k, SM_1)$ and if condition (7.2) above holds, then*

$$x_n + y_n \rightarrow x + y \quad \text{in } D([0, T], \mathbb{R}^k, SM_1).$$

Proof. Apply the characterization of SM_1 convergence in Theorem 12.5.1 (v). At points t in $Disc(x)^c \cap Disc(y)^c$, use the local uniform convergence in Lemma 12.5.1 and Corollary 12.11.1. For other t not in $Disc(x) \cap Disc(y)$, use Theorem 12.6.1. For $t \in Disc(x) \cap Disc(y)$, exploit condition (7.2) to deduce that, for all $\epsilon > 0$, there exists δ and n_0 such that

$$w_s(x_n + y_n, t, \delta) \leq w_s(x_n, t, \delta) + w_s(y_n, t, \delta) + \epsilon$$

for all $n \geq n_0$. ■

In Sections (2.2.7)–(2.2.13) of Skorohod (1956), convenient characterizations of convergence in each topology are given for real-valued functions. We can apply Theorem 12.7.2 to develop associated characterizations for $\mathbb{R}^k$ -valued functions. For each $x \in D([0, T], \mathbb{R}^1)$, $0 \leq t_1 < t_2 \leq T$ and, for each $a < b$ in $\mathbb{R}$, let $v_{t_1, t_2}^{a, b}(x)$ be the number of visits to the strip $[a, b]$ on the interval $[t_1, t_2]$; i.e., $v_{t_1, t_2}^{a, b}(x) = k$ if it is possible to find k (but not $k + 1$) points t'_i such that $t_1 < t'_1 < \dots < t'_k \leq t_2$ such that either

$$x(t_1) \in [a, b], \quad x(t'_1) \notin [a, b], \quad x(t'_2) \in [a, b], \dots,$$

or

$$x(t_1) \notin [a, b], \quad x(t'_1) \in [a, b], \quad x(t'_2) \notin [a, b], \dots$$

We say that $x \in D([0, T], \mathbb{R})$ has a *local maximum (minimum) value at t relative to (t_1, t_2)* in $(0, T)$ if $t_1 < t < t_2$ and either

$$(i) \quad \sup\{x(s) : t_1 \leq s \leq t_2\} \leq x(t) \quad (\inf\{x(s) : t_1 \leq s \leq t_2\} \geq x(t))$$

or

$$(ii) \quad \sup\{x(s) : t_1 \leq s \leq t_2\} \leq x(t-) \quad (\inf\{x(s) : t_1 \leq s \leq t_2\} \geq x(t-)) .$$

We say that x has a *local maximum (minimum) value at t* if it has a local maximum (minimum) value at t relative to some interval (t_1, t_2) with $t_1 < t < t_2$. We call local maximum and minimum values *local extreme values*.

Lemma 12.7.1. (local extreme values) *Any $x \in D([0, T], \mathbb{R})$ has at most countably many local extreme values.*

If b is not a local extreme value of x , then x crosses level b whenever x hits b ; i.e., if b is not a local extreme value and if $x(t) = b$ or $x(t-) = b$, then for every t_1, t_2 with $t_1 < t < t_2$ there exist t'_1, t'_2 with $t_1 < t'_1, t'_2 < t_2$ such that $x(t'_1) < b$ and $x(t'_2) > b$. This property implies the following lemma.

Lemma 12.7.2. *Consider an interval $[t_1, t_2]$ with $0 < t_1 < t_2 < T$. If $x(t_i) \notin \{a, b\}$ for $i = 1, 2$ and a, b are not local extreme values of x , then x crosses one of the levels a and b at each of the $v_{t_1, t_2}^{a, b}(x)$ visits to the strip $[a, b]$ in $[t_1, t_2]$.*

Theorem 12.7.4. (characterization of SM_1 convergence in terms of convergence of number of visits to strips) *There is convergence $d_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ in $D([0, T], \mathbb{R}^k)$ if and only if*

$$v_{t_1, t_2}^{a, b}(\eta x_n) \rightarrow v_{t_1, t_2}^{a, b}(\eta x) \quad \text{as } n \rightarrow \infty$$

for all $\eta \in \mathbb{R}^k$, all points $t_1, t_2 \in \{T\} \cup \text{Disc}(x)^c$ with $t_1 < t_2$ and almost all a, b with respect to Lebesgue measure.

12.8. Topological Completeness

In this section we exhibit a complete metric topologically equivalent to the incomplete metric d_s in (3.7) inducing the SM_1 topology. Since a product metric defined as in (3.11) inherits the completeness of the component metrics, we also succeed in constructing complete metrics inducing the associated product topology. We make no use of the complete metrics beyond showing that the topology is topologically complete. Another approach to topological completeness would be to show that D is homeomorphic to a G_δ subset of a complete metric space, as noted in Section 11.2.

In our construction of complete metrics, we follow the argument used by Prohorov (1956, Appendix 1) to show that the J_1 topology is topologically complete; we incorporate an oscillation function into the metric. For M_1 , we use $w_s(x, \delta)$ in (5.1). Since $w_s(x, \delta) \rightarrow 0$ as $\delta \rightarrow 0$ for each $x \in D$, we need to appropriately “inflate” differences for small δ . For this purpose, let

$$\hat{w}_s(x, z) \equiv \begin{cases} w_s(x, e^z), & z < 0 \\ w_s(x, 1), & z \geq 1. \end{cases} \quad (8.1)$$

Since $w_s(x, \delta)$ is nondecreasing in δ , $\hat{w}_s(x, z)$ is nondecreasing in z . Note that $\hat{w}_s(x, z)$ as a function of z has the form of a cumulative distribution function (cdf) of a finite measure. On such cdf’s, the Lévy metric λ is known to be a complete metric inducing the topology of pointwise convergence at all continuity points of the limit; i.e.,

$$\lambda(F_1, F_2) \equiv \inf\{\epsilon > 0 : F_2(x - \epsilon) - \epsilon \leq F_1(x) \leq F_2(x + \epsilon) + \epsilon\}. \quad (8.2)$$

The Helly selection theorem, p. 267 of Feller (1971), can be used to show that the metric λ is complete.

Thus, our new metric is

$$\hat{d}_s(x_1, x_2) \equiv d_s(x_1, x_2) + \lambda(\hat{w}_s(x_1, \cdot), \hat{w}_s(x_2, \cdot)) . \quad (8.3)$$

Theorem 12.8.1. (a complete SM_1 metric) *The metric $\hat{d}_s$ on D in (8.3) is complete and topologically equivalent to d_s .*

Example 12.8.1. *The counterexample for d_s is not fundamental under $\hat{d}_s$.* Recall that Example 12.10.1 was used to show that the metric d_s is not complete. That example has $x_n = I_{[1, 1+n^{-1}]}$, so that $d_s(x_m, x_n) \rightarrow 0$ as $m, n \rightarrow \infty$, i.e., the sequence $\{x_n\}$ is fundamental for d_s even though it does not converge. Note that $w_s(x_n, \delta) = 1$ for $\delta > 1/2n$ and $w_s(x_n, \delta) = 0$ otherwise. Hence, $\hat{w}_s(x_n, z) = 1$ for $z > \log(1/2n) = -\log(2n)$ and $\hat{w}_s(x_n, z) = 0$ otherwise. Note that $\hat{w}_s(x_n, \cdot)$ corresponds to the cdf of a unit point mass at $-\log(2n)$. Consequently, $\hat{d}_s(x_m, x_n) \not\rightarrow 0$ as $m, n \rightarrow \infty$.

Remark 12.8.1. *An alternative complete metric.* An alternative complete metric topologically equivalent to d_s is

$$d_s^\dagger(x_1, x_2) = m_s(x_1, x_2) + \lambda(\hat{w}_s(x_1, \cdot), \hat{w}_s(x_2, \cdot)) , \quad (8.4)$$

where $m_s \equiv d_{M_2}$ is the M_2 metric in (5.4) of Section 11.5. That is actually what Prohorov did for J_1 (with $\hat{w}_s$ in (8.4) replaced by the J_1 oscillation function). ■

12.9. Non-Compact Domains

It is often convenient to consider the function space $D([0, \infty), \mathbb{R}^k)$ with domain $[0, \infty)$ instead of $[0, T]$. More generally, we may consider the function space $D(I, \mathbb{R}^k)$, where I is a subinterval of the real line. Common cases besides $[0, \infty)$ are $(0, \infty)$ and $(-\infty, \infty) \equiv \mathbb{R}$.

Given the function space $D(I, \mathbb{R}^k)$ for any subinterval I , we define convergence $x_n \rightarrow x$ with some topology to be convergence in $D([a, b], \mathbb{R}^k)$ with that same topology for the restrictions of x_n and x to the compact interval $[a, b]$ for all points a and b that are elements of I and either boundary points of I or are continuity points of the limit function x . For example, for $I = [c, d)$ with $-\infty < c < d < \infty$, we include $a = c$ but exclude $b = d$; for $I = [c, d]$, we include both c and d .

For simplicity, we henceforth consider only the special case in which $I = [0, \infty)$. In that setting, we can equivalently define convergence $x_n \rightarrow x$ as $n \rightarrow \infty$ in $D([0, \infty), \mathbb{R}^k)$ with some topology to be convergence $x_n \rightarrow x$ as $n \rightarrow \infty$ in $D([0, t], \mathbb{R}^k)$ with that topology for the restrictions of x_n and x to $[0, t]$ for $t = t_k$ for each t_k in some sequence $\{t_k\}$ with $t_k \rightarrow \infty$ as $k \rightarrow \infty$, where $\{t_k\}$ can depend on x . It suffices to let t_k be continuity points of the limit function x ; for the J_1 topology, see Stone (1963), Lindvall (1973), Whitt (1980) and Jacod and Shiryaev (1987). We will discuss only the SM_1 topology here, but the discussion applies to the other non-uniform topologies as well. We also will omit most proofs.

As a first step, we consider the case of closed bounded intervals $[t_1, t_2]$. The space $D([t_1, t_2], \mathbb{R}^k)$ is essentially the same as (homeomorphic to) the space $D([0, T], \mathbb{R}^k)$ already studied, but we want to look at the behavior as we change the interval $[t_1, t_2]$. For $[t_3, t_4] \subseteq [t_1, t_2]$, we consider the restriction of x in $D([t_1, t_2], \mathbb{R}^k)$ to $[t_3, t_4]$, defined by

$$r_{t_3, t_4} : D([t_1, t_2], \mathbb{R}^k) \rightarrow D([t_3, t_4], \mathbb{R}^k)$$

with $r_{t_3, t_4}(x)(t) = x(t)$ for $t_3 \leq t \leq t_4$. Let d_{t_1, t_2} be the metric d_s on $D([t_1, t_2], \mathbb{R}^k)$. We want to relate the distance $d_{t_1, t_2}(x_1, x_2)$ and convergence $d_{t_1, t_2}(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ for different domains. We first state a result enabling us to go from the domains $[t_1, t_2]$ and $[t_2, t_3]$ to $[t_1, t_3]$ when $t_1 < t_2 < t_3$.

Lemma 12.9.1. (metric bounds) *For $0 \leq t_1 < t_2 < t_3$ and $x_1, x_2 \in D([t_1, t_3], \mathbb{R}^k)$,*

$$d_{t_1, t_3}(x_1, x_2) \leq d_{t_1, t_2}(x_1, x_2) \vee d_{t_2, t_3}(x_1, x_2) .$$

We now observe that there is an equivalence of convergence provided that the internal boundary point is a continuity point of the limit function.

Lemma 12.9.2. *For $0 \leq t_1 < t_2 < t_3$ and $x, x_n \in D([t_1, t_3], \mathbb{R}^k)$, with $t_2 \in \text{Disc}(x)^c$, $d_{t_1, t_3}(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ if and only if $d_{t_1, t_2}(x_n, x) \rightarrow 0$ and $d_{t_2, t_3}(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$.*

For $x \in D([0, T], \mathbb{R}^k)$ and $0 \leq t_1 < t_2 \leq T$, let $r_{t_1, t_2} : D([0, T], \mathbb{R}^k) \rightarrow D([t_1, t_2], \mathbb{R}^k)$ be the restriction map, defined by $r_{t_1, t_2}(x)(s) = x(s)$, $t_1 \leq s \leq t_2$.

Corollary 12.9.1. (continuity of restriction maps) *If $x_n \rightarrow x$ as $n \rightarrow \infty$ in $D([0, T], \mathbb{R}^k, SM_1)$ and if $t_1, t_2 \in \text{Disc}(x)^c$, then*

$$r_{t_1, t_2}(x_n) \rightarrow r_{t_1, t_2}(x) \quad \text{as } n \rightarrow \infty \quad \text{in } D([t_1, t_2], \mathbb{R}^k, SM_1) .$$

Let $r_t : D([0, \infty), \mathbb{R}^k) \rightarrow D([0, t], \mathbb{R}^k)$ be the *restriction map* with $r_t(x)(s) = x(s)$, $0 \leq s \leq t$. Suppose that $f : D([0, \infty), \mathbb{R}^k) \rightarrow D([0, \infty), \mathbb{R}^k)$ and $f_t : D([0, t], \mathbb{R}^k) \rightarrow D([0, t], \mathbb{R}^k)$ for $t > 0$ are functions with

$$f_t(r_t(x)) = r_t(f(x))$$

for all $x \in D([0, \infty), \mathbb{R}^k)$ and all $t > 0$. We then call the functions f_t *restrictions of the function f* .

Theorem 12.9.1. (continuity from continuous restrictions) *Suppose that $f : D([0, \infty), \mathbb{R}^k) \rightarrow D([0, \infty), \mathbb{R}^l)$ has continuous restrictions f_t with some topology for all $t > 0$. Then f itself is continuous in that topology.*

We now consider the extension of Lipschitz properties to subsets of $D([0, \infty), \mathbb{R}^k)$. For this purpose, suppose that μ_t is one of the M_1 metrics on $D([0, t], \mathbb{R}^k)$ for $t > 0$. As in Section 2 of Whitt(1980), an associated metric μ can be defined on $D([0, \infty), \mathbb{R}^k)$ by

$$\mu(x_1, x_2) = \int_0^\infty e^{-t} [\mu_t(r_t(x_1), r_t(x_2)) \wedge 1] dt. \quad (9.1)$$

The following result implies that the integral in (9.1) is well defined.

Theorem 12.9.2. (regularity of the metric $\mu_t(x_1, x_2)$ as a function of t) *Let μ_t be one of the M_1 metrics on $D([0, t], \mathbb{R}^k)$. For all $x_1, x_2 \in D([0, \infty), \mathbb{R}^k)$, $\mu_t(x_1, x_2)$ as a function of t is right-continuous with left limits in $(0, \infty)$ and has a right limit at 0. Moreover, $\mu_t(x_1, x_2)$ is continuous at $t > 0$ whenever x_1 and x_2 are both continuous at t .*

We also have the following result, paralleling Lemma 2.2 and Theorem 2.5 of Whitt (1980). For (iii), we exploit Theorem 12.5.1 (i).

Theorem 12.9.3. (characterizations of SM_1 convergence with domain $[0, \infty)$) *Suppose that μ and μ_t , $t > 0$ are the SM_1 (or WM_1) metrics on $D([0, \infty), \mathbb{R}^k)$ and $D([0, t], \mathbb{R}^k)$. Then the following are equivalent for x and x_n , $n \geq 1$, in $D([0, \infty), \mathbb{R}^k)$.*

- (i) $\mu(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$;
- (ii) $\mu_t(r_t(x_n), r_t(x)) \rightarrow 0$ as $n \rightarrow \infty$ for all $t \notin \text{Disc}(x)$;

(iii) there exist parametric representations (u, r) and (u_n, r_n) of x and x_n mapping $[0, \infty)$ into the graphs such that

$$\|u_n - u\|_t \vee \|r_n - r\|_t \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

for each $t > 0$.

We now show that the Lipschitz property extends from $D([0, t], \mathbb{R}^k)$ to $D([0, \infty), \mathbb{R}^k)$.

Theorem 12.9.4. (functions with Lipschitz restrictions are Lipschitz) *If a function*

$$f : D([0, \infty), \mathbb{R}^k) \rightarrow D([0, \infty), \mathbb{R}^k)$$

has restrictions

$$f_t : D([0, T], \mathbb{R}^k) \rightarrow D([0, T], \mathbb{R}^k)$$

satisfying

$$\mu_t^2(f_t(r_t(x_1)), f_t(r_t(x_2))) \leq K \mu_t^1(r_t(x_1), r_t(x_2)) \quad \text{for all } t > 0,$$

where K is independent of t , then

$$\mu^2(f(x_1), f(x_2)) \leq (K \vee 1) \mu^1(x_1, x_2).$$

Proof. By (9.1) and the conditions,

$$\begin{aligned} \mu^2(f(x_1), f(x_2)) &= \int_0^\infty e^{-t} [\mu_t^2(r_t(f(x_1)), r_t(f(x_2))) \wedge 1] dt \\ &= \int_0^\infty e^{-t} [\mu_t^2(f_t(r_t(x_1)), f_t(r_t(x_2))) \wedge 1] dt \\ &\leq \int_0^\infty e^{-t} [K \mu_t^1(r_t(x_1), r_t(x_2)) \wedge 1] dt \\ &\leq (K \vee 1) \int_0^\infty e^{-t} [\mu_t^1(r_t(x_1), r_t(x_2)) \wedge 1] dt \\ &\leq (K \vee 1) \mu^1(x_1, x_2). \quad \blacksquare \end{aligned}$$

12.10. Strong and Weak M_2 Topologies

We now define strong and weak versions of Skorohod's M_2 topology. In Section 12.11 we will show that it is possible to define the M_2 topologies by a minor modification of the definitions in Section 12.3, in particular, by

simply using parametric representations in which only r is nondecreasing instead of (u, r) , but now we will use Skorohod's (1956) original approach, and relate it to the Hausdorff metric on the space of graphs.

The weak topology will be defined just like the strong, except it will use the thick graphs G_x instead of the thin graphs Γ_x . In particular, let

$$\mu_s(x_1, x_2) \equiv \sup_{(z_1, t_1) \in \Gamma_{x_1}} \inf_{(z_2, t_2) \in \Gamma_{x_2}} \{ \|(z_1, t_1) - (z_2, t_2)\| \} \quad (10.1)$$

and

$$\mu_w(x_1, x_2) \equiv \sup_{(z_1, t_1) \in G_{x_1}} \inf_{(z_2, t_2) \in G_{x_2}} \{ \|(z_1, t_1) - (z_2, t_2)\| \} . \quad (10.2)$$

Following Skorohod (1956), we say that $x_n \rightarrow x$ as $n \rightarrow \infty$ for a sequence or net $\{x_n\}$ in the strong M_2 topology, denoted by SM_2 if $\mu_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$. Paralleling that, we say that $x_n \rightarrow x$ as $n \rightarrow \infty$ in the weak M_2 topology, denoted by WM_2 , if $\mu_w(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$. We say that $x_n \rightarrow x$ as $n \rightarrow \infty$ in the product topology if $\mu_s(x_n^i, x^i) \rightarrow 0$ (or equivalently $\mu_w(x_n^i, x^i) \rightarrow 0$) as $n \rightarrow \infty$ for each i , $1 \leq i \leq k$.

We can also generate the SM_2 and WM_2 topologies using the Hausdorff metric in (5.2) of Section 11.5. As in (5.4) in Section 11.5, for $x_1, x_2 \in D$,

$$m_s(x_1, x_2) \equiv m_H(\Gamma_{x_1}, \Gamma_{x_2}) = \mu_s(x_1, x_2) \vee \mu_s(x_2, x_1) , \quad (10.3)$$

$$m_w(x_1, x_2) \equiv m_H(G_{x_1}, G_{x_2}) = \mu_w(x_1, x_2) \vee \mu_w(x_2, x_1) \quad (10.4)$$

and

$$m_p(x_1, x_2) \equiv \max_{1 \leq i \leq k} m_s(x_1^i, x_2^i) . \quad (10.5)$$

We will show that the metric m_s induces the SM_2 topology.

That will imply that the metric m_p induces the associated product topology. However, it turns out that the metric m_w does *not* induce the WM_2 topology. We will show that the WM_2 topology coincides with the product topology, so that the Hausdorff metric can be used to define the WM_2 topology via m_p in (10.5).

Closely paralleling the d or M_1 metrics, we have $m_p \leq m_s$ on $D([0, T], \mathbb{R}^k)$ and $m_p = m_w = m_s$ on $D([0, T], \mathbb{R}^1)$. Just as with d , we use m without subscript when the functions are real valued. Example 12.3.1, which showed that WM_1 is strictly weaker than SM_1 also shows that WM_2 is strictly weaker than SM_2 . Example 12.3.3 shows that the SM_2 topology is strictly weaker than the SM_1 topology.

Note that μ_s in (10.1) is *not* symmetric in its two arguments. We first show that if $\mu_s(x, x_n) \rightarrow 0$ as $n \rightarrow \infty$, we need not have $\mu_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$.

Example 12.10.1. *Lack of symmetry of μ_s in its arguments.* To see that we can have $\mu_s(x, x_n) \rightarrow 0$ as $n \rightarrow \infty$ without $\mu_w(x_n, x) \rightarrow 0$ or $\mu_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, let $x(t) = 0$, $0 \leq t \leq 2$, and let $x_n = I_{[1, 1+n^{-1})}$ in $D([0, 2], \mathbb{R}^1)$. Clearly $m_w(x_n, x) \not\rightarrow 0$, but for any $(0, t) \in \Gamma_x = G_x$, we can find $(0, t_n) \in \Gamma_{x_n} = G_{x_n}$ such that $|t_n - t| \rightarrow 0$. ■

We now observe that m_s induces the SM_2 topology.

Theorem 12.10.1. (the Hausdorff metric m_s induces the SM_2 topology)
If $\mu_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, then $\mu_s(x, x_n) \rightarrow 0$ as $n \rightarrow \infty$. Hence, $\mu_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ if and only if $m_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$.

It may seem natural to consider a weak M_2 topology defined by the metric $m_w(x_1, x_2)$ in (10.4), but this does not yield a desirable topology.

Example 12.10.2. *Deficiency of the m_w metric.* To see a deficiency of the m_w metric in (10.4), we show that convergence $d_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, which implies $m_s(x_n, x) \rightarrow 0$, does not imply $\mu_w(x, x_n) \rightarrow 0$ or $m_w(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$. For this purpose, consider x and x_n , $n \geq 1$, in $D([0, 2], \mathbb{R}^2)$ defined by $x^1 = x^2 = I_{[1, 2]}$ and $x_n^1(t) = x_n^2(t) = n(t-1)I_{[1, 1+n^{-1})}(t) + I_{[1+n^{-1}, 2]}(t)$ for $n \geq 1$. Then $d_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ and thus $m_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, but the thick ranges of the graphs of x and x_n are $\rho(G_x) = [0, 2] \times [0, 2]$ and $\rho(G_{x_n}) = \{\alpha(0, 0) + (1-\alpha)(2, 2) : 0 \leq \alpha \leq 1\}$, so that $\mu_w(x, x_n) \not\rightarrow 0$ and $m_w(x_n, x) \not\rightarrow 0$ as $n \rightarrow \infty$. in this case, $\mu_w(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$.

We now observe that m_p induces the WM_2 topology.

Theorem 12.10.2. (WM_2 is the product topology) $\mu_w(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ for μ_w in (10.2) if and only if $m_p(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ for m_p in (10.5), so that the WM_2 topology on $D([0, T], \mathbb{R}^k)$ coincides with the product topology.

We conclude this section by summarizing the relations among the primary distances under consideration in the following theorem.

Theorem 12.10.3. (comparison of distances) *For each $x_1, x_2 \in D$,*

$$d_p \leq d_w \leq d_s \leq d_{J_1} \leq \|\cdot\| ,$$

$$m_p \leq d_p \quad \text{and} \quad m_p \leq m_s \leq d_s .$$

Remark 12.10.1. *Relating the J and M topologies.* The J_i topologies were related to the M_i topologies in a revealing way in Pomarede (1976). The J_2 topology is induced by the Hausdorff metric on the space of incomplete graphs; that shows that J_2 is stronger than M_2 . Similarly, the J_1 topology can be defined in terms of a metric applied to parametric representations of the incomplete graphs; that shows that J_1 is stronger than M_1 . ■

12.11. Alternative Characterizations of M_2 Convergence

We now give alternative characterizations of the SM_2 and WM_2 topologies.

12.11.1. M_2 Parametric Representations

We first observe that the SM_2 and WM_2 topologies can be defined just like the SM_1 and WM_1 topologies in Section 12.3. For this purpose, we say that a *strong M_2 (SM_2) parametric representation* of x is a continuous function (u, r) mapping $[0, 1]$ onto Γ_x such that r is nondecreasing. A *weak M_2 (WM_2) parametric representation* of x is a continuous function mapping $[0, 1]$ into G_x such that r is nondecreasing with $r(0) = 0$, $r(1) = T$ and $u(1) = x(T)$. The corresponding M_1 parametric representations are nondecreasing using the order defined on the graphs Γ_x and G_x in Section 2. In contrast, only the component function r is nondecreasing in the M_2 parametric representations. Let $\Pi_{s,2}(x)$ and $\Pi_{w,2}(x)$ be the sets of all SM_2 and WM_2 parametric representations of x .

Paralleling (3.7) and (3.8), define the distance functions

$$d_{s,2}(x_1, x_2) \equiv \inf_{\substack{(u_j, r_j) \in \Pi_{s,2}(x_j) \\ j=1,2}} \{ \|u_1 - u_2\| \vee \|r_1 - r_2\| \} \quad (11.1)$$

and

$$d_{w,2}(x_1, x_2) \equiv \inf_{\substack{(u_j, r_j) \in \Pi_{w,2}(x_j) \\ j=1,2}} \{ \|u_1 - u_2\| \vee \|r_1 - r_2\| \} . \quad (11.2)$$

We then can say that $x_n \rightarrow x$ as $n \rightarrow \infty$ for a sequence or net $\{x_n\}$ if $d_{s,2}(x_n, x) \rightarrow 0$ or $d_{w,2}(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$. A difficulty with this

approach, just as for the WM_1 topology, is that neither $d_{s,2}$ nor $d_{w,2}$ is a metric.

Example 12.11.1. *Neither $d_{s,2}$ nor $d_{w,2}$ is a metric.* To see that neither $d_{s,2}$ nor $d_{w,2}$ is a metric, consider real-valued functions, so that $d_{s,2} = d_{w,2} = d_2$. Let $x = 2I_{[1,2]}$, $x_{2n+1} = 2I_{[1-2n^{-1}, 1-n^{-1}]} + 2I_{[1,2]}$ and $x_{2n} = I_{[1-n^{-1}, 1]} + 2I_{[1,2]}$ in $D([0, 2], \mathbb{R})$ for $n \geq 3$. For each n , it is possible to choose parametric representations of x_n and x such that $d_2(x_{2n+1}, x) \leq 2n^{-1}$ and $d_2(x_{2n}, x) \leq n^{-1}$. However, $d_2(x_{2n}, x_{2n+1}) \geq 1$ for all n . We cannot simultaneously match the points in $\{2\} \times [1 - 2n^{-1}, 1 - n^{-1}] \subseteq \Gamma_{x_{2n+1}}$ to $\{2\} \times [1, 2] \subseteq \Gamma_{x_{2n}}$ and the points in $\{0\} \times [(1 - n^{-1}), 1] \subseteq \Gamma_{x_{2n+1}}$ to $\{0\} \times [0, 1 - n^{-1}] \subseteq \Gamma_{x_{2n}}$ because the times are inconsistently ordered. ■

12.11.2. SM_2 Convergence

We now establish the equivalence of several alternative characterizations of convergence in the SM_2 topology. To have a characterization involving the local behavior of the functions, we use the uniform-distance function $\bar{w}_s(x, x_2, t, \delta)$ in (4.6). We also use the related uniform-distance functions

$$\bar{w}_s(x_1, x_2, \delta) \equiv \sup_{0 \leq t \leq T} \bar{w}(x_1, x_2, t, \delta) . \quad (11.3)$$

$$\bar{w}_s^*(x_1, x_2, t, \delta) \equiv \|x_1(t) - [x_2((t - \delta) \vee 0), x_2((t + \delta) \wedge T)]\| \quad (11.4)$$

$$\bar{w}_s^*(x_1, x_2, \delta) \equiv \sup_{0 \leq t \leq T} \bar{w}_s^*(x_1, x_2, t, \delta) . \quad (11.5)$$

We now define new oscillation functions. The first is

$$\bar{w}_s^*(x, t, \delta) \equiv \sup\{\|x(t) - [x(t_1), x(t_2)]\|\} , \quad (11.6)$$

where the supremum is over

$$0 \vee (t - \delta) \leq t_1 \leq [0 \vee (t - \delta)] + \delta/2 \text{ and } [T \wedge (t + \delta)] - \delta/2 \leq t_2 \leq (t + \delta) \wedge T.$$

The second is

$$\bar{w}_s^*(x, \delta) \equiv \sup_{0 \leq t \leq T} \bar{w}_s^*(x, t, \delta) . \quad (11.7)$$

The uniform-distance function $\bar{w}_s^*(x_1, x_2, \delta)$ in (11.5) and the oscillation function $\bar{w}_s^*(x, \delta)$ in (11.7) were originally used by Skorohod (1956).

As before, T need not be a continuity point of x in $D([0, T], \mathbb{R}^k)$. Unlike for the M_1 topology, we can have $x_n \rightarrow x$ in (D, M_2) without having $x_n(T) \rightarrow x(T)$.

Example 12.11.2. M_2 convergence does not imply pointwise convergence at the right endpoint. To see that M_2 convergence does not imply that $x_n(T) \rightarrow x(T)$, let $x(0 = x(T-) = 0$, $x(T) = 1$,

$$x_n(0) = x_n(T - 2n^{-1}) = x_n(T) = 0$$

and $x_n(T - n^{-1}) = 1$ for $n \geq 1$ with x and x_n defined by linear interpolation elsewhere. It is easy to see that $x_n \rightarrow x$ (M_2), but $x_n(T) \not\rightarrow x(T)$.

Let $v(x, A)$ represent the oscillation of x over the set A as in (2.5).

Theorem 12.11.1. (characterizations of SM_2 convergence) *The following are equivalent characterizations of $x_n \rightarrow x$ as $n \rightarrow \infty$ in (D, SM_2) :*

- (i) $d_{s,2}(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ for $d_{s,2}$ in (11.1); i.e., for any $\epsilon > 0$ and n sufficiently large, there exist $(u, r) \in \Pi_{s,2}(x)$ and $(u_n, r_n) \in \Pi_{s,2}(x_n)$ such that $\|u_n - u\| \vee \|r_n - r\| < \epsilon$.
- (ii) $m_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ for the metric m_s in (10.3).
- (iii) $\mu_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ for μ_s in (10.1).
- (iv) Given $\bar{w}_s(x_1, x_2, \delta)$ defined in (11.3),

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{w}_s(x_n, x, \delta) = 0 .$$

- (v) For each t , $0 \leq t \leq T$,

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{w}_s(x_n, x, t, \delta) = 0$$

for $\bar{w}_s(x_1, x_2, t, \delta)$ in (4.6).

- (vi) For all $\epsilon > 0$ and all n sufficiently large, there exist finite ordered subsets A of Γ_x and A_n of Γ_{x_n} , as in (3.9) where $(z_1, t_1) \leq (z_2, t_2)$ if $t_1 \leq t_2$, of the same cardinality such that $\hat{d}(A, \Gamma_x) < \epsilon$, $\hat{d}(A_n, \Gamma_{x_n}) < \epsilon$ and $d^*(A, A_n) < \epsilon$ for $\hat{d}$ in (3.10) and d^* in (5.6).

- (vii) Given $\bar{w}_s^*(x_1, x_2, \delta)$ defined in (11.5),

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{w}_s^*(x_n, x, \delta) = 0 .$$

(viii) $x_n(t) \rightarrow x(t)$ as $n \rightarrow \infty$ for each t in a dense subset of $[0, T]$ including 0 and

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{w}_s^*(x_n, \delta) = 0$$

for $\bar{w}_s^*(x, \delta)$ in (11.7).

Remark 12.11.1. The equivalence (iii) $\leftrightarrow$ (vii) $\leftrightarrow$ (viii) was established by Skorohod (1956). ■

Remark 12.11.2. There is no analog to characterization (v) involving $\bar{w}_s^*(x_n, x, t, \delta)$ in (11.4) instead of $\bar{w}_s(x_n, x, t, \delta)$. For $t \in \text{Disc}(x)^c$,

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{w}_s^*(x_n, x, t, \delta) = 0$$

implies pointwise convergence $x_n(t) \rightarrow x(t)$, but not the local uniform convergence in Theorem 12.4.1. ■

12.11.3. WM_2 Convergence

Corresponding characterizations of WM_2 convergence follow from Theorem 12.11.1 because the WM_2 topology is the same as the product topology, by Theorem 12.10.2. Let

$$\bar{w}_w(x_1, x_2, \delta) \equiv \sup_{0 \leq t \leq T} \bar{w}_w(x_1, x_2, t, \delta) \quad (11.8)$$

for $\bar{w}_w(x_1, x_2, t, \delta)$ in (4.7).

Theorem 12.11.2. (characterizations of WM_2 convergence) *The following are equivalent characterizations of $x_n \rightarrow x$ as $n \rightarrow \infty$ in (D, WM_2) :*

(i) $d_{w,2}(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ for $d_{w,2}$ in (11.2); i.e., for any $\epsilon > 0$ and all n sufficiently large, there exist $(u, r) \in \Pi_{w,2}(x)$ and $(u_n, r_n) \in \Pi_{w,2}(x_n)$ such that $\|u_n - u\| \vee \|r_n - r\| < \epsilon$.

(ii) $m_p(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ for the metric m_p in (10.5).

(iii) Given $\bar{w}_w(x_1, x_2, \delta)$ defined in (11.8),

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{w}_w(x_n, x, \delta) = 0 .$$

(iv) For each t , $0 \leq t \leq T$,

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{w}_w(x_n, x, t, \delta) = 0 .$$

(v) For all $\epsilon > 0$ and all sufficiently large n , there exist finite ordered subsets A of G_x and A_n of Γ_{x_n} , of common cardinality m as in (3.9) with $(z_1, t_1) \leq (z_2, t_2)$ if $t_1 \leq t_2$, such that $\hat{d}(A, G_x) < \epsilon$, $\hat{d}(A_n, \Gamma_{x_n}) < \epsilon$ and $d^*(A, A_n) < \epsilon$ for all $n \geq n_0$, for $\hat{d}$ in (5.13) and d^* in (5.6).

Theorem 12.11.2 and Section 12.4 show that all forms of M convergence imply uniform convergence to continuous limit functions.

Corollary 12.11.1. (from WM_2 convergence to uniform convergence) Suppose that $m_p(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$.

(i) If $t \in \text{Disc}(x)^c$, then

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} v(x_n, x, t, \delta) = 0 .$$

(ii) If $x \in C$, then $\lim_{n \rightarrow \infty} \|x_n - x\| = 0$.

Proof. For (i) combine Theorems 12.4.1 and 12.11.2. For (ii) add Lemma 12.4.2. ■

Convergence in WM_2 has the advantage that jumps in the converging functions must be inherited by the limit function.

Corollary 12.11.2. (inheritance of jumps) If $x_n \rightarrow x$ in (D, WM_2) , $t_n \rightarrow t$ in $[0, T]$ and $x_n^i(t_n) - x_n^i(t_n-) \geq c > 0$ for all n , then $x^i(t) - x^i(t-) \geq c$.

Proof. Apply Theorem 12.11.2 (iv). ■

Let $J(x)$ be the maximum magnitude (absolute value) of the jumps of the function x in D . We apply Corollary 12.11.2 to show that J is upper semicontinuous.

Corollary 12.11.3. (upper semicontinuity of J) If $x_n \rightarrow x$ in (D, M_2) , then

$$\overline{\lim}_{n \rightarrow \infty} J(x_n) \leq J(x) .$$

Proof. Suppose that $x_n \rightarrow x$ in (D, WM_2) and there exists a subsequence $\{x_{n_k}\}$ such that $J(x_{n_k}) \rightarrow c$. Then there exist further subsubsequences $\{x_{n_{k_j}}\}$ and $\{t_{n_{k_j}}\}$, and a coordinate i , such that $t_{n_{k_j}} \rightarrow t$ for some $t \in [0, T]$ and $|x_{n_{k_j}}^i(t_{n_{k_j}}) - x_{n_{k_j}}^i(t_{n_{k_j}}-)| \rightarrow c$. Then Corollary 12.11.2 implies that $|x^i(t) - x^i(t-)| \geq c$. ■

12.11.4. Additional Properties of M_2 Convergence

We conclude this section by discussing additional properties of the M_2 topologies. First we note that there are direct M_2 analogs of the M_1 results in Theorems 12.6.1, 12.7.1, 12.7.2 and 12.7.3.

Theorem 12.11.3. (extending SM_2 convergence to product spaces) *Suppose that $m_s(x_n, x) \rightarrow 0$ in $D([0, T], \mathbb{R}^k)$ and $m_s(y_n, y) \rightarrow 0$ in $D([0, T], \mathbb{R}^l)$ as $n \rightarrow \infty$. If*

$$Disc(x) \cap Disc(y) = \phi,$$

then

$$m_s((x_n, y_n), (x, y)) \rightarrow 0 \text{ in } D([0, T], \mathbb{R}^{k+l}) \text{ as } n \rightarrow \infty.$$

Corollary 12.11.4. (from WM_2 convergence to SM_2 convergence when the limit is in D_1) *If $m_p(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ and $x \in D_1$, then $m_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$.*

Theorem 12.11.4. (Lipschitz property of linear functions of the coordinate functions) *For any $x_1, x_2 \in D([0, T], \mathbb{R}^k)$ and $\eta \in \mathbb{R}^k$,*

$$m(\eta x_1, \eta x_2) \leq (\|\eta\| \vee 1) m_s(x_1, x_2).$$

We have an analog of Corollary 12.7.1 for the M_2 topology.

Corollary 12.11.5. (SM_2 -continuity of addition) *If $m_s(x_n, x) \rightarrow 0$ and $m_s(y_n, y) \rightarrow 0$ in $D([0, T], \mathbb{R}^k)$ and*

$$Disc(x) \cap Disc(y) = \phi,$$

then

$$m_s(x_n + y_n, x + y) \rightarrow 0 \text{ in } D([0, T], \mathbb{R}^k).$$

Theorem 12.11.5. (characterization of SM_2 convergence by convergence of all linear functions of the coordinates) *There is convergence $x_n \rightarrow x$ in $D([0, T], \mathbb{R}^k)$ as $n \rightarrow \infty$ in the SM_2 topology if and only if $\eta x_n \rightarrow \eta x$ in $D([0, T], \mathbb{R}^1)$ as $n \rightarrow \infty$ in the M_2 topology for all $\eta \in \mathbb{R}^k$.*

Just as with the M_1 topology, we can get convergence of sums under more general conditions than in Corollary 12.11.5. It suffices to have the jumps of x^i and y^i have common sign for all i . We can express this property by the condition (7.2).

Theorem 12.11.6. (continuity of addition at limits with jumps of common sign) *If $x_n \rightarrow x$ and $y_n \rightarrow y$ in $D([0, T], \mathbb{R}^k, SM_2)$ and if condition (7.2) holds, then*

$$x_n + y_n \rightarrow x + y \quad \text{in} \quad D([0, T], \mathbb{R}^k, SM_2) .$$

We now apply Theorem 12.11.5 to extend a characterization of convergence due to Skorohod (1956) to $\mathbb{R}^k$ -valued functions. For each $x \in D([0, T], \mathbb{R}^1)$ and $0 \leq t_1 < t_2 \leq T$, let

$$M_{t_1, t_2}(x) \equiv \sup_{t_1 \leq t \leq t_2} x(t) . \quad (11.9)$$

The proof exploits the SM_2 analog of Corollary 12.9.1.

Theorem 12.11.7. (characterization of SM_2 convergence in terms of convergence of local extrema) *There is convergence $m_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ in $D([0, T], \mathbb{R}^k)$ if and only if*

$$M_{t_1, t_2}(\eta x_n) \rightarrow M_{t_1, t_2}(\eta x) \quad \text{as} \quad n \rightarrow \infty$$

for all $\eta \in \mathbb{R}^k$ and all points $t_1, t_2 \in \{T\} \cup \text{Disc}(x)^c$ with $t_1 < t_2$.

We can apply the characterization of M_2 convergence in Theorem 12.11.7 to show the preservation of convergence under bounding functions in the M_2 topology.

Corollary 12.11.6. (preservation of WM_2 convergence within bounding functions) *Suppose that*

$$y_n^i(t) \leq x_n^i(t) \leq z_n^i(t)$$

for all $t \in [0, T]$, $1 \leq i \leq k$, and all n . If $m_p(y_n, x) \rightarrow 0$ and $m_p(z_n, x) \rightarrow 0$ as $n \rightarrow \infty$, then $m_p(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$.

Example 12.11.3. *Failure with other topologies.* To see that there is no analog of Corollary 12.11.6 for the M_1 and J_1 topologies, for $n \geq 1$, let $x = I_{[1, 2]}$, $y_n = I_{[1+n^{-1}, 2]}$, $z_n = I_{[1-n^{-1}, 2]}$,

$$x_n(0) = x_n(1 - n^{-1}) = x_n(1 - (3n)^{-1}) = x_n(1 - (5n)^{-1}) = 0$$

and

$$x_n(1 - (2n)^{-1}) = x_n(1 - (4n)^{-1}) = x_n(1) = x_n(2) = 1 ,$$

with x_n defined by linear interpolation elsewhere. Then $y_n(t) \leq x_n(t) \leq z_n(t)$ for all t and n , $y_n \rightarrow x$ and $z_n \rightarrow x$ as $n \rightarrow \infty$ in $D([0, 2], \mathbb{R})$ with the J_1 topology, while $x_n \rightarrow x$ with the M_2 topology, but not with the M_1 , J_2 and J_1 topologies.

12.12. Compactness

We now characterize compact subsets in $D \equiv D([0, T], \mathbb{R}^k)$ in the M topologies, closely following Section 2.7 of Skorohod (1956). To do so, we define new oscillation functions that include more control of the behavior of the functions at the interval endpoints 0 and T . First let

$$\bar{w}_w^*(x, \delta) \equiv \max_{1 \leq i \leq k} \bar{w}_s^*(x^i, \delta) \quad (12.1)$$

for $\bar{w}_s^*$ in (11.7). Given $w_s(x, \delta)$ in (5.1), $w_w(x, \delta)$ in (5.12), $\bar{w}_s^*(x, \delta)$ in (11.7), $\bar{w}_w^*(x, \delta)$ in (12.1) and $\bar{v}(x, t, \delta)$ in (4.3), let

$$w'_s(x, \delta) \equiv w_s(x, \delta) \vee \bar{v}(x, 0, \delta) \vee \bar{v}(x, T, \delta), \quad (12.2)$$

$$w'_w(x, \delta) \equiv w_w(x, \delta) \vee \bar{v}(x, 0, \delta) \vee \bar{v}(x, T, \delta), \quad (12.3)$$

$$\bar{w}'_s(x, \delta) \equiv \bar{w}_s^*(x, \delta) \vee \bar{v}(x, 0, \delta) \vee \bar{v}(x, T, \delta), \quad (12.4)$$

$$\bar{w}'_w(x, \delta) \equiv \bar{w}_w^*(x, \delta) \vee \bar{v}(x, 0, \delta) \vee \bar{v}(x, T, \delta). \quad (12.5)$$

Since

$$\bar{w}_w^*(x, \delta) \leq \bar{w}_s^*(x, \delta) \quad \text{and} \quad \bar{w}_w^*(x, \delta) \leq w_w(x, \delta) \leq w_s(x, \delta)$$

for all $x \in D$ and $\delta > 0$,

$$\bar{w}'_w(x, \delta) \leq \bar{w}'_s(x, \delta) \quad \text{and} \quad \bar{w}'_w(x, \delta) \leq w'_w(x, \delta) \leq w'_s(x, \delta)$$

for all $x \in D$ and $\delta > 0$.

We start by stating a characterization of WM_2 convergence. The proof draws on Theorem 12.11.1.

Theorem 12.12.1. (another characterization of WM_2 convergence) *If $\{x_n\}$ is a sequence in D such that $x_n(t)$ converges as $n \rightarrow \infty$ for all t in a dense subset of $[0, T]$ including 0 and T and*

$$\lim_{\delta \downarrow 0} \overline{\lim_{n \rightarrow \infty}} \bar{w}'_w(x_n, \delta) = 0 \quad (12.6)$$

for $\bar{w}'$ in (12.5), then there exists $x \in D$ such that $m_p(x_n, x) \rightarrow 0$.

Example 12.12.1. *Need for the $\bar{v}$ terms.* To see the need for the terms $\bar{v}(x, 0, \delta)$ and $\bar{v}(x, T, \delta)$ in $\bar{w}'_w(x, \delta)$, let $x_n(0) = 1$, $x_n(n^{-1}) = x_n(1) = 0$ with x_n defined by linear interpolation elsewhere on $[0, 1]$. Then $\bar{w}_s^*(x_n, \delta) = 0$ for all n and δ , but $\{x_n : n \geq 1\}$ does not converge and is not compact in $D([0, 1], \mathbb{R}, M_2)$. Since $\sup_n \bar{v}(x_n, 0, \delta) = 1$ for all $\delta > 0$, (12.6) fails.

Corollary 12.12.1. (new characterizations of convergence in other topologies) *If the conditions of Theorem 12.12.1 hold with $\bar{w}'_w$ in (12.5) replaced by $\bar{w}'_s$ in (12.4), w'_w in (12.3) or w'_s in (12.2), then the convergence can be strengthened to SM_2 , WM_1 or SM_1 , respectively.*

Theorem 12.12.2. (characterizations of compactness) *A subset A of D has compact closure in the SM_1 , WM_1 , SM_2 or WM_2 topology if*

$$\sup_{x \in A} \{\|x\|\} < \infty \quad (12.7)$$

and

$$\limsup_{\delta \downarrow 0} \sup_{x \in A} \{w'(x, \delta)\} < \infty, \quad (12.8)$$

where w' is w'_s in (12.2) for SM_1 , w'_w in (12.3) for WM_1 , $\bar{w}'_s$ in (12.4) for SM_2 and $\bar{w}'_w$ in (12.5) for SM_2 . The conditions are necessary for SM_1 and WM_1 .

Example 12.12.2. *The conditions are not necessary for M_2 .* To see that the conditions in Theorem 12.12.2 are not necessary for the M_2 topologies, for $s \in [1/4, 1/2]$, let

$$x_s = I_{[s, 1/4+s/2]} + I_{[1/2, 1]}$$

in $D([0, 1], \mathbb{R})$. The set $\{x_s : 1/4 \leq s \leq 1/2\}$ is clearly M_2 compact, but

$$\sup_{1/4 \leq s \leq 1/2} \bar{w}_w(x_s, \delta) = 1$$

for all δ , $0 < \delta < 1/4$. ■

Compactness characterizations on D translate into tightness characterizations for sets of probability measures on D . Recall from Chapter 11 that a set A of probability measures on a metric space (S, m) is said to be tight if for all $\epsilon > 0$ there exists a compact subset K of (S, m) such that

$$P(K) > 1 - \epsilon \quad \text{for all } P \in A.$$

Theorem 12.12.3. (characterizations of tightness) *A sequence $\{P_n : n \geq 1\}$ of probability measures on D with the SM_1 , WM_1 , SM_2 or WM_2 topology is tight if the following two conditions hold:*

(i) For each $\epsilon > 0$, there exists c such that

$$P_n(\{x \in D : \|x\| > c\}) < \epsilon, \quad n \geq 1 .$$

(ii) For each $\epsilon > 0$ and $\eta > 0$, there exists $\delta > 0$ such that

$$P_n(\{x \in D : w'(x, \delta) \geq \eta\}) < \epsilon, \quad n \geq 1 ,$$

for w' being the appropriate oscillation function in (12.2)–(12.5). The conditions are also necessary for the SM_1 and WM_1 topologies.

Proof. Suppose that conditions (i) and (ii) hold, where w' is w'_s in (12.2) for SM_1 , w'_w in (12.3) for WM_1 , $\bar{w}'_s$ in (12.4) for SM_2 and $\bar{w}'_w$ in (12.5) for WM_2 . For $\epsilon > 0$ given, choose c and δ_k such that $P_n(A_k^c) < \epsilon 2^{-(k+1)}$, $k \geq 0$, where

$$A_0 = \{x \in D : \|x\| \leq c\} \tag{12.9}$$

and

$$A_k = \{x \in D : w'(x, \delta_k) < k^{-1}\}, \quad k \geq 1 . \tag{12.10}$$

Then let $A = \cap_{k \geq 0} A_k$. By the construction,

$$P_n(A^c) = P_n(\cup_{k \geq 0} A_k^c) \leq \sum_{k=0}^{\infty} P_n(A_k^c) \leq \epsilon . \tag{12.11}$$

Since $A \subseteq A_0$ and

$$\limsup_{\delta \downarrow 0} w'(x, \delta) = 0 , \tag{12.12}$$

the set A has compact closure by Theorem 12.12.1. Going the other way, assume that the topology is SM_1 or WM_1 and suppose that $\{P_n : n \geq 1\}$ is tight, so that for any $\epsilon > 0$ there exists a compact subset K of D such that $P_n(K) > 1 - \epsilon$. By Theorem 12.12.2, for any $\eta > 0$ given, $K \subseteq \{x : \|x\| \leq c\}$ for some c and $K \subseteq \{x : w'(x, \delta) \leq \eta\}$ for small enough δ ; by the monotonicity of $w'(x, \delta)$ in δ for the SM_1 and WM_1 topologies. Hence conditions (i) and (ii) hold for all n . ■

For an alternative characterization of M_1 tightness in $D([0, T], \mathbb{R})$, see Avram and Taqqu (1989).

Chapter 13

Useful Functions

13.1. Introduction

In this chapter we consider several useful functions from D or $D \times D$ to D that can be exploited to establish new stochastic-process limits from given ones. We concentrate on four basic functions introduced in Section 3.5: composition, supremum, reflection and inverse. Another basic function is addition, but it has already been treated in Sections 12.6, 12.7 and 12.11. Our treatment of useful functions follows Whitt (1980), but the emphasis there was on the J_1 topology, even though the M_1 topology was used in places. In contrast, here the emphasis is on the M_1 and M_2 topologies, although we also give results for the J_1 topology. As in the last chapter, many proofs are omitted. Most of the missing proofs appear in Chapter 7 of the Internet Supplement.

Here is how this chapter is organized: We start in Section 13.2 by considering the composition map, which plays an important role in establishing FCLTs involving a random time change. We consider composition without centering in Section 13.2; then we consider composition with centering in Section 13.3.

In Section 13.4 we study the supremum function, both with and without centering. In Section 13.5 we apply the supremum results to treat the (one-sided one-dimensional) reflection map, which arises in queueing applications. We study the two-sided reflection map in Section 14.8.

We start studying the inverse function in Section 13.6. We study the inverse map without centering in Section 13.6 and with centering in Section 13.7. In Section 13.8 we apply the results for inverse functions to obtain corresponding results for closely related counting functions.

Application of these convergence-preservation results to stochastic-process

limits are described in Sections 7.3 and 7.4. Section 7.3 contains FCLT's for counting processes, while Section 7.4 contains FCLT's for renewal-reward processes. When there are heavy-tailed distributions, the M_1 topology plays an important role.

In Chapter 3 of the Internet Supplement we discuss pointwise convergence and its preservation under mappings. The perservation of pointwise convergence focuses on relations for individual sample paths, as in the queueing book by El-Taha and Stidham (1999). From Chapter 3 of the Internet Supplement, we see that a function-space setting is not required for all convergence preservation.

13.2. Composition

This section is devoted to the composition function, mapping (x, y) into $x \circ y$, where

$$(x \circ y)(t) \equiv x(y(t)) \quad \text{for all } t.$$

We have in mind a map from $D^k \times D$ into D^k , where $D^k \equiv D([0, \infty), \mathbb{R}^k)$. The situation is much easier when we consider single times and the map is from $D^k \times \mathbb{R}_+$ to $\mathbb{R}^k$. We can still take advantage of the Skorohod topology on D , though. The following is an elementary, but important, consequence of the local uniform convergence established in Section 12.4.

Proposition 13.2.1. (local uniform convergence) *If*

$$(x_n, t_n) \rightarrow (x, t) \quad \text{in } (D^k, WM_2) \times \mathbb{R}_+,$$

where $t \in \text{Disc}(x)^c$, then

$$x_n(t_n) \rightarrow x(t) \quad \text{in } \mathbb{R}^k.$$

We now consider the composition map as a map from $D^k \times D$ to D , where we allow the domains of x and y to be $\mathbb{R}_+ \equiv [0, \infty)$ and we restrict the range of y to be $\mathbb{R}_+$. However, that is not enough; we need additional regularity conditions to have $x \circ y \in D$.

Example 13.2.1. *The need for a condition on y .* To see that $x \circ y$ need not be in D without additional conditions on y , let $x = I_{[2^{-1}, \infty)}$ and $y = 2^{-1} + \sum_{n=1}^{\infty} (-2)^{-n} I_{[2^{-1}-2^{-n}, 2^{-1}-2^{-(n+1)})}$. Then $x, y \in D$, but $x \circ y$ has no limit from the left at $t = 1/2$. ■

Henceforth in this chapter, unless stipulated otherwise, when $D \equiv D^k$, so that the range of functions is $\mathbb{R}^k$, we let D be endowed with the strong version of the J_1 , M_1 or M_2 topology, and simply write J_1 , M_1 or M_2 . It will be evident that most results also hold with the corresponding weaker product topology.

To ensure that $x \circ y \in D$, we will assume that y is also nondecreasing. We begin by defining subsets of $D \equiv D^k \equiv D([0, \infty), \mathbb{R}^k)$ that we will consider. Let D_0 be the subset of all $x \in D$ with $x^i(0) \geq 0$ for all i . Let $D_\uparrow$ and $D_{\uparrow\uparrow}$ be the subsets of functions in D_0 that are nondecreasing and strictly increasing in each coordinate. Let D_m be the subset of functions x in D_0 for which the coordinate functions x^i are monotone (either increasing or decreasing) for each i . Let C_0 , $C_\uparrow$, $C_{\uparrow\uparrow}$ and C_m be the corresponding subsets of C ; i.e., $C_0 \equiv C \cap D_0$, $C_\uparrow \equiv C \cap D_\uparrow$, $C_{\uparrow\uparrow} = C \cap D_{\uparrow\uparrow}$, and $C_m = C \cap D_m$.

It is important that all of these subsets are measurable subsets of D with the Borel σ -fields associated with the non-uniform Skorohod topologies, which all coincide with the Kolmogorov σ -field generated by the projection maps; see Theorems 11.5.2 and 11.5.3.

Lemma 13.2.1. (Measurability of C in D) *C is a closed subset of (D, J_1) and so a measurable (but not closed) subset of D with the M_1 and M_2 topologies.*

Recall that a subset of a topological space is a G_δ subset if it is a countably intersection of open subsets. Clearly, a G_δ subset belongs to the Borel σ -field.

Lemma 13.2.2. (measurability of subsets of C) *C_m is a closed subset of C , $C_\uparrow$ is a closed subset of C_m and $C_{\uparrow\uparrow}$ is a G_δ subset of $C_\uparrow$.*

Proof. For the third relation, note that

$$C_{\uparrow\uparrow} = \bigcap_{p \in Q} \bigcap_{\substack{q \in Q \\ q > p}} \bigcap_{i=1}^k \{x \in C : x^i(q) - x^i(p) > 0\}$$

where Q is the set of rationals in $\mathbb{R}_+$. ■

Lemma 13.2.3. (measurability of subsets of D) *With any of the non-uniform Skorohod topologies, D_0 is a closed subset of D , D_m is a closed subset of D_0 , $D_\uparrow$ is a closed subset of D_m and $D_{\uparrow\uparrow}$ is a G_δ subset of $D_\uparrow$.*

Proof. For the last relation, let $\{t_j\}$ be a countable dense subset of $\mathbb{R}_+$. For each (j, l) , let

$$D_{i,j,l} = \{x \in D_\uparrow : x^i \text{ is constant over } [t_j \wedge t_l, t_j \vee t_l]\} .$$

Then $D_{i,j,l}$ is a closed subset of $D_\uparrow$ and

$$D_{\uparrow\uparrow} = \bigcap_{j=1}^{\infty} \bigcap_{l=1}^{\infty} \bigcap_{i=1}^k (D_\uparrow - D_{i,j,l}) ,$$

so that $D_{\uparrow\uparrow}$ is indeed a G_δ subset of $D_\uparrow$. ■

We now return to the composition map in (12.2), stating the condition for $x \circ y \in D$ as a lemma.

Lemma 13.2.4. (criterion for $x \circ y$ to be in D) *For each $x \in D([0, \infty), \mathbb{R}^k)$ and $y \in D_\uparrow([0, \infty), \mathbb{R}_+)$, $x \circ y \in D([0, \infty), \mathbb{R}^k)$.*

A basic result, from pp. 145, 232 of Billingsley (1968), is the following. The continuity part involves the topology of uniform convergence on compact intervals.

Theorem 13.2.1. (continuity of composition at continuous limits) *The composition map from $D^k \times D_\uparrow^1$ to D^k is measurable and continuous at $(x, y) \in C^k \times C_\uparrow^1$.*

Example 13.2.2. *Composition is not continuous everywhere.* To see that the composition on $D^1 \times D_\uparrow^1$ is not continuous in any of the Skorohod topologies, let $x_n = x = I_{[1/2, 1]}$, $n \geq 1$, $y(t) = 2^{-1}$ and $y_n(t) = 2^{-1} - n^{-1}$, $0 \leq t \leq 1$. Then $x_n = x$ and $\|y_n - y\| = n^{-1} \rightarrow 0$, but $(x_n \circ y_n)(t) = 0$ and $(x \circ y)(t) = 1$, $0 \leq t \leq 1$. ■

Our goal now is to obtain additional positive continuity results under extra conditions. We use the following elementary lemma.

Lemma 13.2.5. *If $y(t) \in \text{Disc}(x)$ and y is strictly increasing and continuous at t , then $t \in \text{Disc}(x \circ y)$.*

Example 13.2.3. *The need for y to be strictly increasing.* To see the need for the condition that y be strictly increasing at t in Lemma 13.2.5, let $x = I_{[1, \infty)}$ and $y(t) = 1$, $t \geq 0$. Then $(x \circ y)(t) = 1$ for all t , so that $x \circ y$ is continuous. Moreover, if $x_n = x$ and $y_n(t) = 1 - n^{-1}$, $t \geq 0$, $n \geq 1$, then $(x_n \circ y_n)(t) = 0$ for all n and t , so that $x_n \circ y_n$ fails to converge to $x \circ y$ for any t . ■

The following is the J_1 result, taken from Whitt (1980). As indicated before, the proof appears in the Internet Supplement. The first J_1 composition results were established by Silvestrov; see Silvestrov (2000) for an account. See Serfozo (1973, 1975) and Gut (1988) for stochastic-process limits involving composition.

Theorem 13.2.2. (J_1 -continuity of composition) *The composition map from $D^k \times D_{\uparrow}^1$ to D^k taking (x, y) into $(x \circ y)$ is continuous at $(x, y) \in (C^k \times D_{\uparrow}^1) \cup (D^k \times C_{\uparrow\uparrow}^1)$ using the J_1 topology throughout.*

We have a different result for the M topologies:

Theorem 13.2.3. (M -continuity of composition) *If $(x_n, y_n) \rightarrow (x, y)$ in $D^k \times D_{\uparrow}^1$ and $(x, y) \in (D^k \times C_{\uparrow\uparrow}^1) \cup (C_m^k \times D_{\uparrow}^1)$, then $x_n \circ y_n \rightarrow x \circ y$ in D^k , where the topology throughout is M_1 or M_2 .*

In most applications we have $(x, y) \in D^k \times C_{\uparrow\uparrow}^1$, as is illustrated by the next section. That part of the M conditions is the same as for J_1 . The mode of convergence in Theorem 13.2.3 for $y_n \rightarrow y$ does not matter, because on $D_{\uparrow}^1$, convergence in the M_1 and M_2 topologies coincides with pointwise convergence on a dense subset of $[0, \infty)$, including 0; see Corollary 12.5.1.

It is easy to see that composition cannot in general yield convergence in a stronger topology, because $x \circ y = x$ and $x_n \circ y_n = x_n$, $n \geq 1$, when $y_n = y = e$, where $e(t) = t$, $t \geq 0$. Unlike for the J_1 topology, the composition map is in general *not* continuous at $(x, y) \in C \times D_{\uparrow}^1$ in the M topologies.

Example 13.2.4. *Why the J_1 and M conditions differ.* To see that composition is not continuous at $(x, y) \in C \times D_{\uparrow}^1$ in the M topologies, let $y, y_n, x = x_n$ be elements of $D([0, \infty), \mathbb{R})$ defined by

$$\begin{aligned} y(0) &= y(.5-) = 0, y(.5) = .25, y(1) = 1, y(t) = t, t > 1, \\ y_n(0) &= y_n(.5 - n^{-1}) = 0, y_n(.5) = .25, y_n(1) = 1, y_n(t) = t, t > 1, \\ x(0) &= x(.25) = x(t) = 0 \quad \text{for } t > 0.25, x(.125) = 1, \end{aligned}$$

with the functions defined by linear interpolation elsewhere. Note that y jumps from 0 to 0.25 at 0.5, while y_n increases from 0 to 0.25 linearly over the interval $[2^{-1} - n^{-1}, 2^{-1}]$ for each n . Hence $y_n \rightarrow y$ in the M topologies but not in the J topologies. Note that $x(y(t)) = 0$, $t \geq 0$, while $x_n(y_n(2^{-1} - (2n)^{-1})) = x_n(.125) = 1$. Hence $x_n \circ y_n \not\rightarrow x \circ y$ as $n \rightarrow \infty$ in any of the Skorohod topologies. ■

We actually prove a more general continuity result, which covers Theorem 13.2.3 as a special case.

Theorem 13.2.4. (more general M -continuity of composition) *Suppose that $(x_n, y_n) \rightarrow (x, y)$ in $D^k \times D_{\uparrow}^1$. If (i) y is continuous and strictly increasing at t whenever $y(t) \in \text{Disc}(x)$ and (ii) x is monotone on $[y(t-), y(t)]$ and $y(t-), y(t) \notin \text{Disc}(x)$ whenever $t \in \text{Disc}(y)$, then $x_n \circ y_n \rightarrow x \circ y$ in D^k , where the topology throughout is M_1 or M_2 .*

Theorem 13.2.3 follows easily from Theorem 13.2.4: First, on $D^k \times C_{\uparrow}^1$, y is continuous, so only condition (i) need be considered; it is satisfied because y is continuous and strictly increasing everywhere. Second on $C_m^k \times D_{\uparrow}^1$, x is continuous so only condition (ii) need be considered; it is satisfied because x is monotone everywhere. Hence it suffices to prove Theorem 13.2.4, which is done in the Internet Supplement. The general idea in our proof of Theorem 13.2.4 is to work with the characterization of convergence using oscillation functions evaluated at single arguments, exploiting Theorems 12.5.1 (v), 12.5.2 (iv), 12.11.1 (v) and 12.11.2 (iv).

13.3. Composition with Centering

We now consider the composition map with centering. To obtain results, we apply both composition and addition. The results yield sufficient conditions for random sums and other processes transformed by a random time change to satisfy FCLTs, as we show in Section 7.4.

We start by establishing convergence properties of composition plus addition. We state results for the J_1 topology as well as the M_1 and M_2 topologies. As before, let e be the identity map on $[0, \infty)$.

Theorem 13.3.1. (convergence preservation for composition plus addition) *Let x, z and $x_n, n \geq 1$ be elements of D^k ; let y, y_n and $v_n, n \geq 1$ be elements of $D_{\uparrow}^1$; and let $c_n \in \mathbb{R}^k$ for $n \geq 1$. If*

$$(x_n - c_n e, y_n, c_n(y_n - v_n)) \rightarrow (x, y, z) \quad \text{in } D^k \times D_{\uparrow}^1 \times D^k, \quad (3.1)$$

$y \in C_{\uparrow\uparrow}^1$ and

$$\text{Disc}(x \circ y) \cap \text{Disc}(z) = \emptyset, \quad (3.2)$$

then

$$x_n \circ y_n - c_n v_n \rightarrow x \circ y + z \quad \text{in } D^k, \quad (3.3)$$

where the topology throughout is J_1 , M_1 or M_2 . If the topology is M_1 or M_2 , then instead of (3.2) it suffices for $x^i \circ y$ and z^i to have no common discontinuities with jumps of the opposite sign for $1 \leq i \leq k$.

Proof. Note that

$$x_n \circ y_n - c_n v_n = (x_n - c_n e) \circ y_n + c_n (y_n - v_n) .$$

For the M topologies, apply Theorem 13.2.3 for composition, using the condition $y \in C_{\uparrow\uparrow}^1$, and Corollaries 12.7.1 and 12.11.5 for addition with the M_1 and M_2 topologies, respectively. The J_1 result is proved similarly, using Theorem 13.2.2 instead of Theorem 13.2.3. For addition with J_1 , use Remark 12.6.2. Use Theorems 12.7.3 and 12.11.6 for the weaker condition for addition to be continuous with the M topologies. ■

The standard application of Theorem 13.3.1 has $c_n^i \rightarrow \infty$ as $n \rightarrow \infty$ for each i and $v_n = b_n e$, where $b_n \rightarrow b$. We describe that case below.

Corollary 13.3.1. (convergence preservation for composition with linear centering) *Let x, z and $x_n, n \geq 1$, be elements of D^k ; let $y_n, n \geq 1$, be elements of $D_{\uparrow}^1$; let $c_n \in \mathbb{R}^k$ and $b_n \in \mathbb{R}^1$ satisfy $|c_n^i| \rightarrow \infty$ for each i and $b_n \rightarrow b$ as $n \rightarrow \infty$. If*

$$(x_n - c_n e, c_n (y_n - b_n e)) \rightarrow (x, z) \quad \text{in} \quad D^k \times D^k \quad (3.4)$$

and

$$Disc(x \circ be) \cap Disc(z) = \emptyset , \quad (3.5)$$

then

$$(x_n \circ y_n - c_n b_n e) \rightarrow x \circ y + z \quad \text{in} \quad D^k , \quad (3.6)$$

where the topology throughout is J_1, M_1 or M_2 . If the topology is M_1 or M_2 , then instead of condition (3.5) it suffices for $x^i \circ be$ and z^i to have no common discontinuities with jumps of opposite sign, $1 \leq i \leq k$. ■

Proof. Since $|c_n^i| \rightarrow \infty$ as $n \rightarrow \infty$ for each i , the limit in (3.4) implies that $\|y_n - b_n e\| \rightarrow 0$ as $n \rightarrow \infty$. Hence $\|y_n - be\| \rightarrow 0$ as $n \rightarrow \infty$ and

$$(x_n - c_n e, y_n, c_n (y_n - be)) \rightarrow (x, y, z) \quad \text{in} \quad D^k \times D_{\uparrow}^1 \times D^k ,$$

where $y = be$. Hence we can apply Theorem 13.3.1 to obtain the desired conclusion. ■

We now consider an application of the convergence-preservation results above to obtain a FCLT involving a random time change. Specifically, we consider an application of Corollary 13.3.1. Let $(X_n(t), Y_n(t)) : t \geq 0$ be

random elements of $D^k \times D_{\uparrow}^1$ for each $n \geq 1$, with one of the topologies under consideration. Let $\mathbf{X}_n$, $\mathbf{Y}_n$ and $\mathbf{Z}_n$ be normalized processes constructed by

$$\begin{aligned}\mathbf{X}_n(t) &\equiv \delta_n^{-1}[X_n(nt) - \mu_n nt], \quad t \geq 0 \\ \mathbf{Y}_n(t) &\equiv \delta_n^{-1}[Y_n(nt) - \lambda_n nt], \quad t \geq 0 \\ \mathbf{Z}_n(t) &\equiv \delta_n^{-1}[(X_n(Y_n(nt)) - \lambda_n \mu_n nt)], \quad t \geq 0.\end{aligned}\tag{3.7}$$

Corollary 13.3.2. (stochastic consequence with linear centering) *Suppose that (X_n, Y_n) is a random element of $D^k \times D_{\uparrow}^1$ for each n . If*

$$(\mathbf{X}_n, \mathbf{Y}_n) \Rightarrow (\mathbf{U}, \mathbf{V}) \quad \text{in } D^k \times D^1 \tag{3.8}$$

with topology J_1 , M_1 or M_2 , for the scaled processes $\mathbf{X}_n$, $\mathbf{Y}_n$ in (3.7) with $\delta_n \rightarrow \infty$, $n\delta_n^{-1} \rightarrow \infty$, $\mu_n \rightarrow \mu$ with $\mu^i \neq 0$ for all i and $\lambda_n \rightarrow \lambda$, and if

$$P(\text{Disc}(\mathbf{U} \circ \lambda \mathbf{e}) \cap \text{Disc}(\mathbf{V}) = \emptyset) = 1, \tag{3.9}$$

then

$$\mathbf{Z}_n \Rightarrow \mathbf{U} \circ \lambda \mathbf{e} + \mu \mathbf{V} \quad \text{in } D^k \tag{3.10}$$

for $\mathbf{Z}_n$ in (3.7) and the same topology. If the topology is M_1 or M_2 , then instead of condition (3.9) it suffices for $\mathbf{U}^i \circ \lambda \mathbf{e}$ and $\mathbf{V}^i$ to almost surely have no common discontinuities with jumps of opposite sign, $1 \leq i \leq k$.

Proof. First, since $\mu_n \rightarrow \mu$ as $n \rightarrow \infty$ in $\mathbb{R}^k$, from condition (3.8) we obtain

$$(\mathbf{X}_n, \mu_n \mathbf{Y}_n) \Rightarrow (\mathbf{U}, \mu \mathbf{V}) \quad \text{in } D^k \times D^k \tag{3.11}$$

from the continuous mapping theorem. Now apply Corollary 13.3.1 with $c_n = n\delta_n^{-1}\mu_n$, $b_n = \lambda_n$,

$$x_n(t) = \delta_n^{-1}X_n(nt) \quad \text{and} \quad y_n(t) = n^{-1}Y_n(nt).$$

By the Skorohod (1956) representation theorem, there exist versions of the processes such that almost surely

$$(x_n - c_n e, c_n(y_n - b_n e)) \rightarrow (x, z) \quad \text{as } n \rightarrow \infty$$

where $x = \mathbf{U}$ and $z = \mu \mathbf{V}$. Corollary 13.3.1 then yields

$$(x_n \circ y_n - c_n b_n e) \rightarrow x \circ y + z \quad \text{as } n \rightarrow \infty \tag{3.12}$$

almost surely in D^k , where $y = \lambda \mathbf{e}$, but the limit process in (3.12) is distributed the same as the limit process in (3.10). The almost sure convergence in (3.12) implies the convergence in distribution in (3.10). ■

A standard application of Corollary 13.3.2 is to random sums. Then, for each $n \geq 1$, $\{X_n(nt) : t \geq 0\}$ corresponds to a sequence of partial sums; i.e.,

$$X_n(nt) = \sum_{j=1}^{\lfloor nt \rfloor} Z_{n,j}, \quad t \geq 0,$$

where $\lfloor x \rfloor$ is the greatest integer less than or equal to x and $\{Z_{n,j} : j \geq 1\}$ is a sequence of random vectors in $\mathbb{R}^k$ for each n . The composition then yields a random sum, i.e.,

$$(x_n \circ y_n)(t) = \delta_n^{-1} X_n(Y_n(nt)) = \delta_n^{-1} \sum_{j=1}^{Y_n(nt)} Z_{n,j},$$

so that the limit (3.10) becomes for a random sum. We consider the special case in which the summands $Z_{n,j}$ come from a single IID sequence and the random index $Y_n(t)$ is a renewal process in Section 7.4.

Another application of Corollary 13.3.2 is to establish stochastic-process limits that imply asymptotic validity of sequential stopping rules in stochastic simulations. The asymptotic validity occurs in the limit as the desired volume of the target confidence set decreases. See Chapter 4 of the Internet Supplement.

We now establish a variant of Theorem 13.3.1 with nonlinear centering terms. In the proof we apply continuity of multiplication, which we now establish. By multiplication of x and y in D , we mean $(xy)(t) \equiv x(t)y(t)$ for all t . For the M topologies, the condition on the behavior at common discontinuities is more stringent for multiplication than for addition because of the way signs multiply.

Example 13.3.1. *The need for stronger conditions.* To see the need for stronger conditions with multiplication, let $x_n \equiv -1 + 2I_{[2^{-1}-n^{-1}, \infty)}$ and let $y_n \equiv y \equiv -1 + 2I_{[2^{-1}, \infty)}$ for $n \geq 2$. Then $x_n \rightarrow y$ in (D, J_1) as $n \rightarrow \infty$, but $x_n y_n = 1 - 2I_{[2^{-1}-n^{-1}, 2^{-1}]}$, which does not converge to $y^2 = 1$ in any of the Skorohod topologies. ■

Theorem 13.3.2. (continuity of multiplication) *Suppose that $x_n \rightarrow x$ and $y_n \rightarrow y$ in $D([0, \infty), \mathbb{R})$ with one of the Skorohod topologies J_1 , M_1 or M_2 . If the topology is J_1 , then assume that $\text{Disc}(x) \cap \text{Disc}(y) = \emptyset$. If the topology is*

M_1 or M_2 , then assume for each $t \in \text{Disc}(x) \cap \text{Disc}(y)$ that $x(t), x(t-), y(t)$ and $y(t-)$ are all nonnegative and $[x(t) - x(t-)][y(t) - y(t-)] \geq 0$. Then $x_n y_n \rightarrow xy$ in $D([0, \infty), \mathbb{R})$ with the same topology, where $(xy)(t) \equiv x(t)y(t)$ for $t \geq 0$.

Proof. For J_1 , we can conclude that $(x_n, y_n) \rightarrow (x, y)$ in D^2 by the J_1 analog of Theorem 12.6.1; see Remark 12.6.2. It is then easy to show that $x_n y_n \rightarrow xy$. Use the fact that $x_n \rightarrow x$ implies that $\sup_n \{\|x_n\|\} < \infty$. For M_1 , apply the characterization in Theorem 12.5.1 (v). For M_2 , apply the characterization in Theorem 12.11.7. ■

Theorem 13.3.3. (convergence preservation for composition with nonlinear centering) *Let $x, x_n \in D^k$, $y, y_n \in D_\uparrow^1$, $y \in C_{\uparrow\uparrow}$, x have a continuous derivative $\dot{x}$ and $c_n \rightarrow \infty$. If*

$$c_n(x_n - x, y_n - y) \rightarrow (u, v) \quad \text{in} \quad D^k \times D^1 \quad (3.13)$$

with one of the topologies J_1 , M_1 or M_2 , where

$$\text{Disc}(u \circ y) \cap \text{Disc}(v) = \emptyset, \quad (3.14)$$

then

$$c_n(x_n \circ y_n - x \circ y) \rightarrow u \circ y + (\dot{x} \circ y)v \quad \text{in} \quad D^k \quad (3.15)$$

with the same topology, where

$$[(\dot{x} \circ y)v](t) \equiv [\dot{x}^1(y(t))v(t), \dots, \dot{x}^k(y(t))v(t)]. \quad (3.16)$$

If the topology is M_1 or M_2 , then instead of condition (3.14) it suffices to have $\dot{x}(t) \geq (\leq) 0$ for all t and the functions $u \circ y$ and v to have no common discontinuities with jumps of opposite (common) sign.

Proof. Note that

$$c_n(x_n \circ y_n - x \circ y) = c_n(x_n - x) \circ y_n + c_n(x \circ y_n - x \circ y),$$

Given condition (3.13), we obtain

$$[c_n(x_n - x), c_n(y_n - y), y_n] \rightarrow [u, v, y] \quad \text{in} \quad D^k \times D^1 \times D^1$$

and then, applying composition, multiplication and addition,

$$[c_n(x_n \circ y_n - x \circ y_n) + (\dot{x} \circ y)c_n(y_n - y)] \rightarrow u \circ y + (\dot{x} \circ y)v$$

by virtue of Theorems 13.2.2, 13.2.3 and 13.3.2 and condition (3.14) (or the alternative M -topology condition). Note that

$$\begin{aligned} & \|c_n(x_n \circ y_n - x \circ y) - c_n(x_n \circ y_n - x \circ y_n) - c_n(\dot{x} \circ y)(y_n - y)\| \\ & \leq \|c_n(x \circ y_n - x \circ y) - c_n(\dot{x} \circ y)(y_n - y)\|. \end{aligned} \quad (3.17)$$

However, the term on the right in (3.17) is asymptotically negligible because

$$c_n(x \circ y_n - x \circ y)(t) = c_n \int_{y(t)}^{y_n(t)} \dot{x}(s) ds$$

and

$$\sup_{0 \leq s \leq t} \left| c_n \int_{y(s)}^{y_n(s)} \dot{x}(u) du - \dot{x}(y(s)) c_n(y_n(s) - y(s)) \right| \rightarrow 0 \text{ as } n \rightarrow \infty,$$

because $\dot{x}$ is uniformly continuous over bounded intervals and $\|y_n - y\|_t \rightarrow 0$ as a consequence of $d(c_n(y_n - y), v) \rightarrow 0$. ■

13.4. Supremum

In this section we consider the supremum function, mapping $D \equiv D([0, T], \mathbb{R})$ into itself according to

$$x^\uparrow(t) = \sup_{0 \leq s \leq t} x(s), \quad 0 \leq t \leq T. \quad (4.1)$$

We are primarily interested in the supremum function because it is closely related to the reflection map, discussed in the next section. Another motivation is extreme-value theory; see Resnick (1987) and Embrechts et al. (1997).

We have already observed that the map from D to $\mathbb{R}$ taking x into $x^\uparrow(t)$ is continuous in the M_2 topology at all $t \in \text{Disc}(x)^c$; that is a consequence of Theorem 12.11.7. Now we consider the map from D to D taking x into the function $x^\uparrow$ in (4.1).

The supremum function can be thought of as the *nondecreasing majorant*: It is easy to see that

$$x^\uparrow = \inf\{y \in D : y \geq x, y \text{ nondecreasing}\},$$

where $y \geq x$ if $y(t) \geq x(t)$ for all t . If $x \in D_0$, then $x^\uparrow \in D_\uparrow$.

It is easy to see that the supremum function is Lipschitz in the uniform norm:

Lemma 13.4.1. (Lipschitz property of the supremum function with the uniform norm) *For any $x_1, x_2 \in D([0, T], \mathbb{R})$,*

$$\|x_1^\uparrow - x_2^\uparrow\| \leq \|x_1 - x_2\| .$$

As consequences of Lemma 13.4.1, we obtain corresponding Lipschitz properties with the J_1 , M_1 and M_2 metrics d_{J_1} , d_s and m_s , here denoted by d_{J_1} , d_{M_1} and d_{M_2} . For the M_1 topology, we use the following result.

Lemma 13.4.2. (inheritance of parametric representations) *For any $x \in D$, if $(u, r) \in \Pi(x)$ ($\Pi_{s,2}(x)$), then $(u^\uparrow, r) \in \Pi(x^\uparrow)$ ($\Pi_{s,2}(x)$).*

Theorem 13.4.1. (Lipschitz property of the supremum function) *For any $x_1, x_2 \in D([0, T], \mathbb{R})$,*

$$\begin{aligned} d_{J_1}(x_1^\uparrow, x_2^\uparrow) &\leq d_{J_1}(x_1, x_2) , \\ d_{M_1}(x_1^\uparrow, x_2^\uparrow) &\leq d_{M_1}(x_1, x_2) , \\ d_{M_2}(x_1^\uparrow, x_2^\uparrow) &\leq d_{M_2}(x_1, x_2) . \end{aligned}$$

Example 13.4.1. *Convergence preservation fails with pointwise convergence.*

It is significant that analogs of Lemma 13.4.1 and Theorem 13.4.1 do not hold for pointwise convergence: Let $x_n = I_{[n^{-1}, 2n^{-1}]}$. Then $x_n(t) \rightarrow 0$ as $n \rightarrow \infty$ for all t , while $x_n^\uparrow(t) \rightarrow 1$ as $n \rightarrow \infty$ for all $t > 0$. ■

On the other hand, there is a pointwise-convergence analog of Theorem 13.4.1 for a single function; see Section 3.3 of the Internet Supplement.

Moreover, the conclusion in Theorem 13.4.1 can be recast in terms of pointwise convergence: Since $x^\uparrow$ is nondecreasing, convergence $x_n^\uparrow \rightarrow x^\uparrow$ in the M topologies is equivalent to pointwise convergence at continuity points of $x^\uparrow$, because on $D_\uparrow$ the M_1 and M_2 topologies coincide with pointwise convergence on a dense subset of $\mathbb{R}_+$ including 0 and T ; see Corollary 12.5.1. Thus the M topologies have not contributed much so far. We obtain more useful convergence-preservation results for the supremum map with the M topologies when we combine supremum with centering. As before, let e be the identity map, i.e., $e(t) = t$, $0 \leq t \leq T$. The proof is in the Internet Supplement.

Theorem 13.4.2. (convergence preservation with the supremum function and centering) *Suppose that $c_n(x_n - e) \rightarrow y$ as $n \rightarrow \infty$ in $D([0, T], \mathbb{R})$ with one of the topologies J_1 , M_1 or M_2 , where $c_n \rightarrow \infty$.*

- (a) If the topology is M_1 or M_2 , then $c_n(x_n^\uparrow - e) \rightarrow y$ in the same topology.
 (b) If the topology is J_1 , then $c_n(x_n^\uparrow - e) \rightarrow y$ if and only if y has no negative jumps.

Example 13.4.2. *Pointwise convergence is not enough.* To see that a pointwise convergent analog of Theorem 13.4.2 does not hold, let $x_n = c_n^{-1}I_{[n^{-1}, 2n^{-1}]} + e$ where $c_n \rightarrow \infty$. Then $c_n(x_n - e)(t) = I_{[n^{-1}, 2n^{-1}]}(t) \rightarrow 0$ as $n \rightarrow \infty$ for all $t > 0$, while $x_n^\uparrow(t) = c_n^{-1} + t$ and $c_n(x_n^\uparrow - e)(t) = 1$ for all n sufficiently large, for $t > 0$. ■

A common case covered by Theorem 13.4.2 is $y \in C$. If $y \in C$, then all modes of convergence in Theorem 13.4.2 reduce to uniform convergence and we have $c_n(x_n^\uparrow - e) \rightarrow y$ whenever $c_n(x_n - e) \rightarrow y$. Since $c_n \rightarrow \infty$, under the conditions of Theorem 13.4.2, $\|x_n - e\| \rightarrow 0$ as $n \rightarrow \infty$. By Theorem 13.4.1, $\|x_n^\uparrow - e\| \rightarrow 0$ as well.

We use the following lemma in the proof of both Theorem 13.4.2 above and Theorem 13.4.3 below.

Lemma 13.4.3. *If $x \in D([0, T], \mathbb{R})$ and x has no negative jumps, then for any $\epsilon > 0$ there is a $\delta > 0$ such that*

$$v^-(x, \delta) \equiv \sup_{\substack{0 \vee (t-\delta) \leq t' \leq t \\ 0 \leq t \leq T}} \{x(t') - x(t)\} < \epsilon. \quad (4.2)$$

We can easily extend Theorem 13.4.2 to cover a case of nonlinear centering. Recall that $\Lambda \equiv \Lambda([0, T])$ is the set of increasing homeomorphisms of $[0, T]$. We use elements of Λ as the centering term.

Corollary 13.4.1. (convergence preservation with the supremum and nonlinear centering) *Suppose that $c_n(x_n - \lambda_n) \rightarrow y$ as in $D([0, T], \mathbb{R})$ with one of the topologies J_1 , M_1 or M_2 , where $\lambda_n \rightarrow \lambda$ with $\lambda, \lambda_n \in \Lambda([0, T])$ and $c_n \rightarrow \infty$.*

- (a) *If the topology is M_1 or M_2 , then $c_n(x_n^\uparrow - \lambda_n) \rightarrow y$ in the same topology.*
 (b) *If the topology is J_1 , then $c_n(x_n^\uparrow - \lambda_n) \rightarrow y$ if and only if y has no negative jumps.*

Proof. Given $c_n(x_n - \lambda_n) \rightarrow y$, we have $c_n(x_n \circ \lambda_n^{-1} - e) \rightarrow y \circ \lambda^{-1}$ by applying Theorems 13.2.2 and 13.2.3. Then Theorem 13.4.2 implies that $c_n(x_n^\uparrow \circ \lambda_n^{-1} - e) \rightarrow y \circ \lambda^{-1}$ with the limit holding J_1 if and only if $y \circ$

λ^{-1} has no negative jumps. Clearly, $y \circ \lambda^{-1}$ has no negative jumps if and only if y does. Finally, apply Theorems 13.2.2 and 13.2.3 again to get $c_n(x_n^\uparrow \circ \lambda_n^{-1} \circ \lambda_n - \lambda_n) \rightarrow y \circ \lambda^{-1} \circ \lambda$, which implies the conclusion because $\lambda_n^{-1} \circ \lambda_n = \lambda^{-1} \circ \lambda = e$. ■

We now obtain joint convergence in the stronger topologies on $D([0, T], \mathbb{R}^2)$ under the condition that the limit function have no negative jumps.

Theorem 13.4.3. (criterion for joint convergence) *Suppose that $c_n(x_n - e) \rightarrow y$ as $n \rightarrow \infty$ in $D([0, T], \mathbb{R})$ with one of the J_1 , M_1 or M_2 topologies, where $c_n \rightarrow \infty$. If, in addition, y has no negative jumps, then*

$$c_n(x_n - e, x_n^\uparrow - e) \rightarrow (y, y) \quad \text{as } n \rightarrow \infty \quad (4.3)$$

in $D([0, T], \mathbb{R}^2)$ with the strong version of the same topology, i.e., with SJ_1 , SM_1 or SM_2 .

Since addition is continuous on D^2 with the strong topologies, we obtain the following corollary.

Corollary 13.4.2. *Under the conditions of Theorem 13.4.3,*

$$\|c_n(x_n^\uparrow - x_n)\| \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Example 13.4.3. *The problem with negative jumps.* To see that Corollary 13.4.2 does not hold and the simple direct argument with parametric representations in the proof of Theorem 13.4.3 does not work for Theorem 13.4.2 when there are negative jumps, let $y = -I_{[1/2, 1]}$, $c_n = n$ and $c_n(x_n - e) = y$, i.e., $x_n = e + n^{-1}y$. First,

$$c_n(x_n^\uparrow - x_n)(1/2) = 1 \quad \text{for all } n \geq 1.$$

We now show what goes wrong with the parametric representations. let $u_n = u$ and $r_n = r$ with

$$u(0) = u(1/3) = 0, \quad u(2/3) = u(1) = -1 \quad (4.4)$$

and

$$r(0) = 0, \quad r(1/3) = 1/2 = r(2/3), \quad r(1) = 1,$$

with u and r defined by linear interpolation elsewhere. Then $(u'_n, r) \in \Pi(c_n(x_n^\uparrow - e))$ for $u'_n = (u + nr)^\uparrow - nr$, so that

$$u'_n(0) = u'_n(1/3) = u'_n(2/3) = 0, \quad u'_n(2/3 + n^{-1}) = u'_n(1) = -1 \quad (4.5)$$

with u'_n defined by linear interpolation elsewhere. From (4.4) and (4.5), we see that $|u'_n(2/3) - u(2/3)| = 1$ for all n . Thus, to get the positive result, different parametric representations are needed for $c_n(x_n^\uparrow - e)$. ■

We next give an elementary result about the supremum function when the centering is in the other direction, so that x_n must be rapidly decreasing. Convergence $x_n^\uparrow(t) \rightarrow x(0)$ as $n \rightarrow \infty$ is to be expected, but that conclusion can not be drawn if the M_2 convergence in the condition is replaced by pointwise convergence.

Theorem 13.4.4. (convergence preservation with the supremum function when the centering is in the other direction) *Suppose that $c_n \rightarrow \infty$ and $x_n + c_n e \rightarrow y$ in $D([0, T], \mathbb{R}, M_2)$. Then*

$$\|x_n^\uparrow - z(y)\| \rightarrow 0 \quad \text{as } n \rightarrow \infty ,$$

where $z(y)(t) \equiv y(0)$, $0 \leq t \leq T$.

Example 13.4.4. M_2 convergence cannot be replaced by pointwise convergence. To see that the M_2 convergence cannot be replaced by pointwise convergence in the condition in Theorem 13.4.4, even to get pointwise convergence in the conclusion, let $x(t) = 0$, $0 \leq t \leq 1$, and $x_n(t) = I_{[n^{-1}, 2n^{-1}]}(t) - t$, $0 \leq t \leq 1$, $n \geq 1$. Then $x_n + e \rightarrow x$ pointwise (and not M_2), but $x_n^\uparrow(t) \rightarrow 1$ as $n \rightarrow \infty$ for all $t > 0$.

13.5. One-Dimensional Reflection

Closely related to the supremum function is the one-dimensional (one-sided) reflection mapping, which we have used to construct queueing processes. Indeed, the reflection mapping can be defined in terms of the supremum mapping as

$$\phi(x) \equiv x + (-x \vee 0)^\uparrow ;$$

i.e.,

$$\phi(x)(t) = x(t) - (\inf\{x(s) : 0 \leq s \leq t\} \wedge 0) , \quad 0 \leq t \leq T , \quad (5.1)$$

as in (2.5) in Section 5.2.

The Lipschitz property for the supremum function with the uniform topology in Lemma 13.4.1 immediately implies a corresponding result for the reflection map ϕ in (5.1).

Lemma 13.5.1. (Lipschitz property with the uniform metric) *For any $x_1, x_2 \in D([0, T], \mathbb{R})$,*

$$\|\phi(x_1) - \phi(x_2)\| \leq 2\|x_1 - x_2\| .$$

Proof. By (5.1),

$$\begin{aligned} \|\phi(x_1) - \phi(x_2)\| &\leq \|x_1 - x_2\| + \|(-x_1 \vee 0)^\uparrow - (-x_2 \vee 0)^\uparrow\| \\ &\leq \|x_1 - x_2\| + \|(-x_1 \vee 0) - (-x_2 \vee 0)\| \leq 2\|x_1 - x_2\|. \quad \blacksquare \end{aligned}$$

Example 13.5.1. *The bound is tight.* To see that the bound in Lemma 13.5.1 is tight, let $x_1(t) = 0$, $0 \leq t \leq 1$, and $x_2 = -I_{[1/3, 1/2]} + I_{[1/2, 1]}$ in $D([0, 1], \mathbb{R})$. Then $\phi(x_1) = x_1$, while $\phi(x_2) = 2I_{[1/2, 1]}$, so that $\|x_1 - x_2\| = 1$ and $\|\phi(x_1) - \phi(x_2)\| = 2$.

Unfortunately, however, the Lipschitz property for the reflection map ϕ with the uniform topology does not even imply continuity in all the Skorohod topologies. In particular, ϕ is not continuous in the M_2 topology.

Example 13.5.2. *Continuity fails in M_2 .* To see that the reflection map ϕ in (5.1) is not continuous in the M_2 topology, let $x = -I_{[1, 2]}$ and

$$x_n(0) = x_n(1 - 3n^{-1}) = x(1 - n^{-1}) = 0$$

and

$$x_n(1 - 2n^{-1}) = x_n(1) = x_n(2) = -1$$

with x_n defined by linear interpolation elsewhere. Then $x_n \rightarrow x$ in $D([0, 2], \mathbb{R})$, but $\phi(x)(t) = 0$, $0 \leq t \leq 2$, and $\phi(x_n)(1 - n^{-1}) = 1$, so that $\phi(x_n) \not\rightarrow \phi(x)$. This example fails to be a counterexample for the M_1 topology because then $x_n \not\rightarrow x$ as $n \rightarrow \infty$. $\blacksquare$

We do obtain positive results with the J_1 and M_1 topologies. As before, let d_{J_1} and d_{M_1} be the metrics in equations (3.2) and (3.4) in Section 3.3.. For the J_1 result, we use the following elementary lemma.

Lemma 13.5.2. *For any $x \in D$ and $\lambda \in \Lambda$,*

$$\phi(x) \circ \lambda = \phi(x \circ \lambda) .$$

For the M_1 result, we use the following lemma. A fundamental difficulty for treating the more general multidimensional reflection map is that Lemma 13.5.3 below does not extend to the multidimensional reflection map; see Chapter 14.

Lemma 13.5.3. (preservation of parametric representations under reflections) *For any $x \in D$, if $(u, r) \in \Pi(x)$, then $(\phi(u), r) \in \Pi(\phi(x))$.*

Proof. First, $(\phi(u), r)$ is continuous since (u, r) is, by Lemma 13.5.1. It suffices to show that $(\phi(u)(s), r(s)) \in \Gamma_{\phi(x)}$ for all s and that $(\phi(u), r)$ is nondecreasing in the order on $\Gamma_{\phi(x)}$. If $t \in \text{Disc}(x^c)$, then by (5.1) $\phi(u)(s) = \phi(x)(t)$ for each s such that $r(s) = t$. It remains to consider $t \in \text{Disc}(x)$. There exists an interval $[a, b] \subseteq [0, 1]$ such that $r(s) = t$ for $s \in [a, b]$, $u(a) = x(t-)$ and $u(b) = x(t)$. Moreover, by (5.1), $\phi(u)(a) = \phi(x)(t-)$ and $\phi(u)(b) = \phi(x)(t)$, with $\phi(u)(s)$ moving continuously and monotonically from $\phi(u)(a)$ to $\phi(u)(b)$ as s increases over $[a, b]$. Hence $(\phi(u)(s), r(s)) \in \Gamma_{\phi(x)}$ for all $s \in [0, 1]$ and $(\phi(u), r)$ is nondecreasing in the order on $\Gamma_{\phi(x)}$. ■

Theorem 13.5.1. (Lipschitz property with the J_1 and M_1 metrics) *For any $x_1, x_2 \in D([0, T], \mathbb{R})$,*

$$d_{J_1}(\phi(x_1), \phi(x_2)) \leq 2d_{J_1}(x_1, x_2)$$

and

$$d_{M_1}(\phi(x_1), \phi(x_2)) \leq 2d_{M_1}(x_1, x_2) ,$$

where ϕ is the reflection map in (5.1).

Proof. First, for the J_1 metric, by Lemmas 13.5.2 and 13.5.1,

$$\begin{aligned} d_{J_1}(\phi(x_1), \phi(x_2)) &= \inf_{\lambda \in \Lambda} \{ \|\phi(x_1) \circ \lambda - \phi(x_2)\| \vee \|\lambda - e\| \} \\ &= \inf_{\lambda \in \Lambda} \{ \|\phi(x_1 \circ \lambda) - \phi(x_2)\| \vee \|\lambda - e\| \} \\ &\leq \inf_{\lambda \in \Lambda} \{ 2\|x_1 \circ \lambda - x_2\| \vee \|\lambda - e\| \} \leq 2d_{J_1}(x_1, x_2) . \end{aligned}$$

Turning to M_1 , we use Lemma 13.5.3 to conclude that $(\phi(u), r) \in \Pi(\phi(x))$ whenever $(u, r) \in \Pi(x)$. Then, by Lemma 13.5.1,

$$\begin{aligned} d_{M_1}(\phi(x_1), \phi(x_2)) &= \inf_{\substack{(u_i, r_i) \in \Pi(\phi(x_i)) \\ i=1,2}} \{ \|u_1 - u_2\| \vee \|r_1 - r_2\| \} \\ &\leq \inf_{\substack{(u_i, r_i) \in \Pi(x_i) \\ i=1,2}} \{ \|\phi(u_1) - \phi(u_2)\| \vee \|r_1 - r_2\| \} \\ &\leq \inf_{\substack{(u_i, r_i) \in \Pi(x_i) \\ i=1,2}} \{ 2\|u_1 - u_2\| \vee \|r_1 - r_2\| \} \leq 2d_{M_1}(x_1, x_2) . \end{aligned}$$

Remark 13.5.1. *The Lipschitz constant.* Example 13.5.1 shows that the bounds in Theorem 13.5.1 are tight; i.e., the Lipschitz constant is 2. ■

Theorem 13.5.1 covers the standard heavy-traffic regime for one single-server queue when $\rho = 1$, where ρ is the traffic intensity. The next result covers the other cases: $\rho < 1$ and $\rho > 1$. We use the following elementary lemma in the easy case of the uniform metric.

Lemma 13.5.4. *Let d be the metric for the U , J_1 , M_1 or M_2 topology. Let $x \vee a : D \rightarrow D$ be defined by*

$$(x \vee a)(t) \equiv x(t) \vee a, \quad 0 \leq t \leq T. \quad (5.2)$$

Then, for any $x_1, x_2 \in D$,

$$d(x_1 \vee a(x_1), x_2 \vee a(x_2)) \leq d(x_1, x_2) .$$

Theorem 13.5.2. (convergence preservation with centering) *Suppose that $x_n - c_n e \rightarrow y$ in $D([0, T], \mathbb{R})$ with the U , J_1 , M_1 or M_2 topology.*

(a) If $c_n \rightarrow +\infty$, then

$$\phi(x_n) - c_n e \rightarrow y + \gamma(y) \quad \text{as } n \rightarrow \infty \quad \text{in } D$$

with the same topology, where

$$\gamma(y)(t) \equiv (-y(0)) \vee 0 = -(y(0) \wedge 0), \quad 0 \leq t \leq T.$$

(b) If $c_n \rightarrow -\infty$, $y(0) \leq 0$ and y has no positive jumps, then

$$\|\phi(x_n) - 0e\| \rightarrow 0 \quad \text{as } n \rightarrow \infty \quad \text{in } D ,$$

where $e(t) = t$, $0 \leq t \leq T$.

Example 13.5.3. *The necessity of the condition on $y(0)$.* To see the need for the condition $y(0) \leq 0$ in Theorem 13.5.2 (b), let $y(t) = 1$, $0 \leq t \leq T$, $c_n = -n$ and $x_n(t) = (c_n e + y)(t) = 1 - nt$ for all t . Then $x_n - c_n e = y$ for all n , but $\phi(x_n)(0) = 1$ and $\phi(x_n)(t) \rightarrow 0$ for all $t > 0$.

13.6. Inverse

We now consider the inverse map, which arises in the study of renewal processes, first passage times and extremal processes; see Billingsley (1968), Gut (1988) and Resnick (1987).

It is convenient to consider the inverse map on the subset D_u of x in $D \equiv D([0, \infty), \mathbb{R})$ that are unbounded above and satisfy $x(0) \geq 0$. For $x \in D_u$, let the inverse of x be

$$x^{-1}(t) = \inf\{s \geq 0 : x(s) > t\}, \quad t \geq 0. \quad (6.1)$$

As before, let D_0 be the subset of x in D with $x(0) \geq 0$, and let $D_\uparrow$ and $D_{\uparrow\uparrow}$ be the subsets of nondecreasing and strictly increasing functions in D_0 . Let $D_{u,\uparrow} \equiv D_u \cap D_\uparrow$ and $D_{u,\uparrow\uparrow} \equiv D_u \cap D_{\uparrow\uparrow}$. Clearly,

$$D_{u,\uparrow\uparrow} \subseteq D_{u,\uparrow} \subseteq D_u \subseteq D_0 .$$

13.6.1. The Standard Topologies

Recall that on $D_\uparrow$ the M_1 and M_2 topologies reduce to pointwise convergence on a dense subset including 0. The following result supplements Lemmas 13.2.1–13.2.3.

Lemma 13.6.1. (measurability of D_u) *Let D have one of the topologies J_1 , M_1 or M_2 . The subset D_u is a G_δ subset of D_0 .*

Proof. Note that

$$D_u = \cap_{n=1}^{\infty} (D_0 - \bar{D}_n) ,$$

where $\bar{D}_n$ is the subset of functions in D_0 bounded above by n . In the non-uniform Skorohod topologies, $\bar{D}_n$ is a closed subset of D_0 , so that D_u is a G_δ subset of D_0 . ■

We begin our study of the inverse function by stating some basic results. Our first result shows that the inverse map is closely related to the supremum.

Lemma 13.6.2. (duality) *For any $x \in D_u$, $x^{-1} \in D_{u,\uparrow}$ and $(x^{-1})^{-1} = x^\uparrow$.*

Corollary 13.6.1. *For any $x \in D_{u,\uparrow}$, $(x^{-1})^{-1} = x$.*

Remark 13.6.1. *The left-continuous inverse.* As part of Lemma 13.6.2, x^{-1} is right-continuous. In some circumstances it is convenient to work instead with the left-continuous inverse

$$x^\leftarrow(t) \equiv \inf\{s \geq 0 : x(s) \geq t\}, \quad t \geq 0 . \quad (6.2)$$

For $x \in D_u$, $x^\leftarrow(t) = x^{-1}(t-)$, $t \geq 0$, with $x^{-1}(0-) \equiv 0$. Note that $x^\leftarrow$ need not be right-continuous at 0. Indeed, $x^\leftarrow(0) > 0 = x^\leftarrow(0)$ if and only if $x^{-1}(0) > 0$. If $x^{-1}(0) = 0$, then the completed graphs of x^{-1} in (6.1) and $x^\leftarrow$ in (6.2) are identical, which implies that many M_1 and M_2 results for x^{-1} apply directly to $x^\leftarrow$ as well under that condition. ■

The left-continuous inverse has an appealing inverse property not shared by the right-continuous inverse:

Lemma 13.6.3. (inverse relation) *For any $x \in D_{u,\uparrow}$ and $t_1, t_2 \geq 0$,*

$$x^\leftarrow(t_1) \leq t_2 \quad \text{if and only if} \quad x(t_2) \geq t_1 . \quad (6.3)$$

Lemma 13.6.4. *For any $x \in D_{u,\uparrow}$,*

$$0 \leq (x \circ x^{-1})(t) - t \leq x(x^{-1}(t)) - x(x^{-1}(t)-), \quad (6.4)$$

$$0 \leq (x^{-1} \circ x)(t) - t \leq x^{-1}(x(t)) - x^{-1}(x(t)-), \quad (6.5)$$

$$0 \leq (x \circ x^{\leftarrow})(t) - t \leq x(x^{\leftarrow}(t)) - x(x^{\leftarrow}(t)-), \quad (6.6)$$

$$0 \leq t - (x^{\leftarrow} \circ x)(t) \leq x^{-1}(x(t)) - x^{\leftarrow}(x(t)), \quad (6.7)$$

where $x(0-)$ is interpreted as 0.

Let $J_t(x)$ be the maximum jump of x over $[0, t]$, i.e.

$$J_t(x) \equiv \sup_{0 \leq s \leq t} \{x(t) - x(t-)\}. \quad (6.8)$$

where again $x(0-) \equiv 0$.

Corollary 13.6.2. *For any $x \in D_{u,\uparrow}$ and $t > 0$,*

$$\|x \circ x^{-1} - e\|_t \leq J_{x^{-1}(t)}(x) \quad (6.9)$$

and

$$\|x^{-1} \circ x - e\|_t \leq J_{x(t)}(x^{-1}), \quad (6.10)$$

for $J_t(x)$ in (6.8).

Lemma 13.6.5. *Suppose that $x \in D_{u,\uparrow}$. Then $x \in D_{u,\uparrow\uparrow}$ if and only if $x^{-1} \in C_{u,\uparrow}$.*

We now consider the inverse together with composition applied to elements of $\Lambda \equiv \Lambda([0, \infty))$, i.e., to homeomorphisms of $\mathbb{R}_+ \equiv [0, \infty)$. For each $\lambda \in \Lambda$, $\lambda(0) = 0$ and there is an inverse λ^{-1} with $\lambda, \lambda^{-1} \in C_{\uparrow\uparrow}$ and $\lambda \circ \lambda^{-1} = \lambda^{-1} \circ \lambda = e$.

Lemma 13.6.6. *If $x \in D_{u,\uparrow}$ and $\lambda_1, \lambda_2 \in \Lambda([0, \infty))$, then*

$$(\lambda_1 \circ x \circ \lambda_2)^{-1} = \lambda_2^{-1} \circ x^{-1} \circ \lambda_1^{-1}.$$

Proof. Note that

$$\begin{aligned} (\lambda_1 \circ x \circ \lambda_2)^{-1}(t) &= \inf\{s \geq 0 : (\lambda_1 \circ x \circ \lambda_2)(s) > t\} \\ &= \inf\{s \geq 0 : (x \circ \lambda_2)(s) > \lambda_1^{-1}(t)\} \\ &= \inf\{\lambda_2^{-1}(s) \geq 0 : x(s) > \lambda^{-1}(t)\} \\ &= (\lambda_2^{-1} \circ x^{-1} \circ \lambda_1^{-1})(t). \quad \blacksquare \end{aligned}$$

We now turn to continuity properties of the inverse map. First we note that the inverse map from (D_u, J_1) to (D_u, J_1) or even from (D_u, U) to (D_u, J_1) is in general not continuous.

Example 13.6.1. *The inverse is not continuous when the range has the J_1 topology.* To see that the inverse map from $(D_{u,\uparrow}, U)$ to $(D_{u,\uparrow}, J_1)$ is not continuous, let $x = 2I_{[0,2]} + eI_{[2,\infty)}$ and

$$x_n = (2 - n^{-1})I_{[0,1)} + (2 + n^{-1})I_{[1,2+n^{-1})} + eI_{[2+n^{-1},\infty)}.$$

Then $\|x_n - x\| = n^{-1} \rightarrow 0$ and $x_n^{-1} \rightarrow x^{-1}(M_1)$, but $x_n^{-1} \not\rightarrow x^{-1}(J_1)$. ■

Even for the M_1 topology, there are complications at the left endpoint of the domain $[0, \infty)$.

Example 13.6.2. *Complications at the left endpoint of the domain.* To see that the inverse map from $(D_{u,\uparrow}, U)$ to $(D_{u,\uparrow}, M_1)$ is in general not continuous, let $x(t) = 0$, $0 \leq t < 1$, and $x(t) = t$, $t \geq 1$; Let $x_n = t/n$, $0 \leq t < 1$ and $x_n(t) = t$, $t \geq 1$. Then $\|x_n - x\|_\infty = n^{-1} \rightarrow 0$, but $x_n^{-1}(0) = 0 \not\rightarrow 1 = x^{-1}(0)$, so that $x_n^{-1} \not\rightarrow x^{-1}(M_1)$. ■

To avoid the problem in Example 13.6.2, we can require that $x^{-1}(0) = 0$. To develop an equivalent condition, let $D_{u,\epsilon}^\uparrow$ be the subset of functions x in D_u such that $x(t) = 0$ for $0 \leq t \leq \epsilon$.

Then let

$$D_u^* \equiv \bigcap_{n=1}^\infty (D_{u,n^{-1}})^\epsilon. \quad (6.11)$$

Lemma 13.6.7. (measurability of D_u^*) *With the J_1 , M_1 or M_2 topology, D_u^* in (6.11) is a G_δ subset of D_u and*

$$D_u^* = \{x \in D_u : x^{-1}(0) = 0\}. \quad (6.12)$$

Let $D_{u,\uparrow}^* \equiv D_{\uparrow} \cap D_u^*$. A key property of $D_{u,\uparrow}^*$, not shared by $D_{u,\uparrow}$ because of the complication at the origin, is that parametric representation (u, r) for x directly serve as parametric representations for x^{-1} when we switch the roles of the components u and r .

Lemma 13.6.8. (switching the roles of u and r) *For $x \in D_{u,\uparrow}^*$, the graph Γ_x serves as the graph of $\Gamma_{x^{-1}}$ with the axes switched. Thus, $(u, r) \in \Pi(x)$ if and only if $(r, u) \in \Pi(x^{-1})$, where $\Pi(x)$ is the set of M_1 parametric representations.*

Corollary 13.6.3. (continuity on (D_u^*, M_1)) *The inverse map from (D_u^*, M_1) to $(D_{u,\uparrow}, M_1)$ is continuous.*

Proof. First apply Theorem 13.4.1 for the supremum. Then apply Lemma 13.6.8. ■

We now generalize Corollary 13.6.3 by only requiring that the limit be in D_u^* . As before, the missing proof is in the Internet Supplement.

Theorem 13.6.1. (measurability and continuity at limits in D_u^*) *The inverse map in (6.1) from (D_u, M_2) to $(D_{u,\uparrow}, M_1)$ is measurable and continuous at $x \in D_u^*$, i.e., for which $x^{-1}(0) = 0$.*

Corollary 13.6.4. . (continuity at strictly increasing functions) *The inverse map from (D_u, M_2) to $(D_{u,\uparrow}, U)$ is continuous at $x \in D_{u,\uparrow}$.*

Proof. First, $D_{u,\uparrow\uparrow} \subseteq D_{u,\uparrow}^*$, so that we can apply Theorem 13.6.1 to get $x_n^{-1} \rightarrow x^{-1}$ in $(D_{u,\uparrow}, M_1)$. However, by Lemma 13.6.4, $x^{-1} \in C$ when $x \in D_{u,\uparrow\uparrow}$. Hence the M_1 convergence $x_n^{-1} \rightarrow x^{-1}$ actually holds in the stronger topology of uniform convergence over compact subsets. ■

13.6.2. The M'_1 Topology

For cases in which the condition $x^{-1}(0) = 0$ in Theorem 13.6.1 is not satisfied, we can modify the M_1 and M_2 topologies to obtain convergence, following Puhalskii and Whitt (1997). With these new weaker topologies, which we call M'_1 and M'_2 , we do not require that $x_n(0) \rightarrow x(0)$ when $x_n \rightarrow x$. We construct the new topologies by extending the graph of each function x by appending the segment $[0, x(0)] \equiv \{\alpha 0 + (1 - \alpha)x(0) : 0 \leq \alpha \leq 1\}$. Let the new graph of $x \in D$ be

$$\begin{aligned} \Gamma'_x = \{ & (z, t) \in \mathbb{R}^k \times [0, \infty) : z = \alpha x(t) + (1 - \alpha)x(t-) \\ & \text{for } 0 \leq \alpha \leq 1 \text{ and } t \geq 0\}, \end{aligned} \quad (6.13)$$

where $x(0-) \equiv 0$. Let $\Pi'(x)$ and $\Pi'_2(x)$ be the sets of all M_1 and M_2 parametric representations of Γ'_x , defined just as before. We say that $x_n \rightarrow x$ in (D, M'_1) if there exist parametric representations $(u_n, r_n) \in \Pi'(x_n)$ and $(u, r) \in \Pi'(x)$, where $\Pi'(x)$ is the set of M'_1 parametric representations of x , such that

$$\|u_n - u\|_t \vee \|r_n - r\|_t \rightarrow 0 \quad \text{as } n \rightarrow \infty \quad \text{for each } t > 0. \quad (6.14)$$

We have a corresponding definition of convergence in (D, M'_2) using the parametric representations in $\Pi'_2(x)$ instead of $\Pi'(x)$. With the M'_i topologies, we obtain a cleaner statement than Lemma 13.6.8.

Lemma 13.6.9. (graphs of the inverse with the M'_i topology) *For $x \in D_{u,\uparrow}$, the graph Γ'_x serves as the graph $\Gamma'_{x^{-1}}$ with the axes switched, so that $(u, r) \in \Pi'(x)$ ($\Pi'_2(x)$) if and only if $(r, u) \in \Pi'(x^{-1})$ ($\Pi'_2(x^{-1})$).*

Thus we get an alternative to Theorem 13.6.1.

Theorem 13.6.2. (continuity in the M'_1 topology) *The inverse map in (6.1) from (D_u, M'_2) to $(D_{u,\uparrow}, M'_1)$ is continuous.*

Proof. By the M'_2 analog of Theorem 13.4.1, if $x_n \rightarrow x$ in (D_u, M'_2) , then $x_n^\uparrow \rightarrow x^\uparrow$ in $(D_{u,\uparrow}, M'_2)$. Since the M'_2 topology coincides with the M'_1 topology on $D_\uparrow$, we get $x_n^\uparrow \rightarrow x^\uparrow$ in $(D_{u,\uparrow}, M'_1)$. By Lemma 13.6.9, we get $(x_n^\uparrow)^{-1} \rightarrow (x^\uparrow)^{-1}$ in $(D_{u,\uparrow}, M'_1)$. That gives the desired result because $(x^\uparrow)^{-1} = x^{-1}$ for all $x \in D_u$. ■

An alternative approach to the difficulty at the origin besides M'_i topology on $D_u([0, \infty), \mathbb{R})$ is the ordinary M_i topology on $D_u((0, \infty), \mathbb{R})$. The difficulty at the origin goes away if we ignore it entirely, which we can do by making the function domain $(0, \infty)$ for the image of the inverse functions.

In particular, Theorem 13.6.2 implies the following corollary.

Corollary 13.6.5. (continuity when the origin is removed from the domain) *The inverse map in (6.1) from $D_u([0, \infty), M_2)$ to $D_{u,\uparrow}((0, \infty), M_1)$ is continuous.*

Proof. Since the M'_2 topology is weaker than M_2 , if $x_n \rightarrow x$ in $D_u([0, \infty), M_2)$, then $x_n \rightarrow x$ in $D_u([0, \infty), M'_2)$. Apply Theorem 13.6.2 to get $x_n^{-1} \rightarrow x^{-1}$ in $D_{u,\uparrow}([0, \infty), M'_1)$. That implies $x_n^{-1} \rightarrow x^{-1}$ for the restrictions in $D_\uparrow([t_1, t_2], M_1)$ for all $t_1, t_2 \in \text{Disc}(x^{-1})^c$, which in turn implies that $x_n^{-1} \rightarrow x^{-1}$ in $D_{u,\uparrow}((0, \infty), M_1)$. ■

However, in general we cannot work with the inverse on $D_u((0, \infty), \mathbb{R})$.

Example 13.6.3. *Difficulty with the domain $(0, \infty)$.* To see the problem with having the function domain be $(0, \infty)$, let $x = e$ and $x_n(0) = x_n(2n^{-1}) = 0$, $x_n(n^{-1}) = 1$, $x_n(t) = t - 2n^{-1}$, $t \geq 2n^{-1}$, with x_n defined by linear interpolation elsewhere. Then $x_n \rightarrow x$ in $D((0, \infty), \mathbb{R}, U)$, but $x_n^{-1} \not\rightarrow x^{-1} \equiv e$, because $x_n^{-1}(t) \rightarrow 1$ as $n \rightarrow \infty$ for each t with $0 < t < 1$. ■

We can obtain positive results if all the functions are required to be monotone. The following result is elementary.

Theorem 13.6.3. (equivalent characterizations of convergence for monotone functions) *For x_n , $n \geq 1$, $x \in D_{u,\uparrow}([0, \infty), \mathbb{R})$, the following are equivalent:*

$$x_n \rightarrow x \quad \text{in} \quad D_{u,\uparrow}((0, \infty), \mathbb{R}, M_1) ; \quad (6.15)$$

$$x_n \rightarrow x \quad \text{in} \quad D_{u,\uparrow}([0, \infty), \mathbb{R}, M'_1) ; \quad (6.16)$$

$$x_n(t) \rightarrow x(t) \quad \text{for all } t \text{ in a dense subset of } (0, \infty) ; \quad (6.17)$$

$$x_n^{-1} \rightarrow x^{-1} \quad \text{in} \quad D((0, \infty), \mathbb{R}, M_1) ; \quad (6.18)$$

$$x_n^{-1} \rightarrow x^{-1} \quad \text{in} \quad D([0, \infty), \mathbb{R}, M'_1) ; \quad (6.19)$$

$$x_n^{-1}(t) \rightarrow x^{-1}(t) \quad \text{for all } t \text{ in a dense subset of } (0, \infty). \quad (6.20)$$

Example 13.6.4. *The need for monotonicity.* To see the advantage of M'_1 on $[0, \infty)$ over M_1 on $(0, \infty)$, let $x(t) = 1$, $t \geq 0$,

$$x_n^1(0) = 0, x_n^1(n^{-1}) = 1 = x_n^1(t), \quad t \geq n^{-1}, \quad (6.21)$$

and

$$x_n^2(0) = 0 = x_n^2(2n^{-1}), x_n^2(n^{-1}) = x_n^2(3n^{-1}) = 1 = x_n^2(t), \quad t \geq 3n^{-1}, \quad (6.22)$$

with x_n^1 and x_n^2 defined by linear interpolation elsewhere. Then $x_n^1 \rightarrow x$ in both $D((0, \infty), \mathbb{R}, M_1)$ and in $D([0, \infty), \mathbb{R}, M'_1)$, but $x_n^2 \rightarrow x$ only in $D([0, \infty), \mathbb{R}, M'_1)$. The monotonicity condition provides the equivalence in Theorem 13.6.3.

13.6.3. First Passage Times

In this final subsection we consider some real-valued functions closely related to the inverse function. Sometimes we are interested in the first passage time to or beyond some specified level. Given any specified level $z \in \mathbb{R}$, the *first passage time* beyond z is the function $\tau_z : D_u \rightarrow \mathbb{R}$ defined in terms of the inverse function by

$$\tau_z(x) \equiv x^{-1}(z). \quad (6.23)$$

It is elementary that τ_z has the following two scaling invariance properties: For any $c > 0$,

$$\tau_{cz}(cx) = \tau_z(x) \quad (6.24)$$

and

$$c\tau_z(x \circ ce) = \tau_z(x), \quad (6.25)$$

where e is the identity map, i.e., $e(t) = t$ for $t \geq 0$.

Three functions closely related to the first-passage-time function τ_z are the *overshoot function* $\gamma_z : D_u \rightarrow \mathbb{R}$ defined by

$$\gamma_z(x) \equiv x(\tau_z(x)) - z , \quad (6.26)$$

the *last-value function* $\lambda_z : D_u \rightarrow \mathbb{R}$ defined by

$$\lambda_z(x) \equiv x(\tau_z(x)-) \quad (6.27)$$

and the *final-jump functions* $\delta_z : D_u \rightarrow \mathbb{R}$ defined by

$$\delta_z(x) \equiv x(\tau_z(x)) - x(\tau_z(x)-) . \quad (6.28)$$

The following continuity properties are elementary, but of course important. It clearly does not suffice to have pointwise convergence.

Theorem 13.6.4. (continuity of first-passage-time functions) *Let x be an element of D_u that is not equal to z throughout the interval $(\tau_z(x) - \epsilon, \tau_z(x))$ for any $\epsilon > 0$. If $x_n \rightarrow x$ in (D, M_2) , then*

$$(\tau_z(x_n), \gamma_z(x_n), \lambda_z(x_n), \delta_z(x_n)) \rightarrow (\tau_z(x), \gamma_z(x), \lambda_z(x), \delta_z(x))$$

as $n \rightarrow \infty$ in $\mathbb{R}^4$.

The regularity condition holds almost surely for Lévy processes. Hence we have the following consequence of Theorem 13.6.4, which we apply to queues in Section 9.7.

Theorem 13.6.5. (convergence of first-passage-time functions for Lévy limit processes) *Let X be a Lévy process such that*

$$P(\overline{\lim}_{t \rightarrow \infty} X(t) = \infty) = 1 . \quad (6.29)$$

If $X_n \Rightarrow X$ in (D_u, M_2) , then

$$(\tau_z(X_n), \gamma_z(X_n), \lambda_z(X_n), \delta_z(X_n)) \rightarrow (\tau_z(X), \gamma_z(X), \lambda_z(X), \delta_z(X))$$

in $\mathbb{R}^4$ for any $z > 0$.

13.7. Inverse with Centering

We continue considering the inverse map, but now with centering. We start by considering linear centering. In particular, we consider when a limit for $c_n(x_n - e)$ implies a limit for $c_n(x_n^{-1} - e)$ when $x_n \in D_u \equiv D_u([0, \infty), \mathbb{R})$ and $c_n \rightarrow \infty$. By considering the behavior at one t , it is natural to anticipate that we should have $c_n(x_n^{-1} - e) \rightarrow -y$ when $c_n(x_n - e) \rightarrow y$. A first step for the M topologies is to apply Theorem 13.4.2, which yields limits for $c_n(x_n^\uparrow - e)$. Thus for the M topologies, it suffices to assume that $x_n \in D_{u,\uparrow}$.

For the J_1 topology, however, a different argument is needed to get limits when $y \notin C$, as the following result shows.

Lemma 13.7.1. *Suppose that $x_n \in D_u$, $n \geq 1$, and $c_n \rightarrow \infty$. if $c_n(x_n^\uparrow - e) \rightarrow y$ and $c_n(x_n^{-1} - e) \rightarrow -y$ (J_1), then $y \in C$.*

Proof. Since $x_n^\uparrow \in D_{u,\uparrow}$, $c_n(x_n^\uparrow - e)$ has no negative jumps. Since the topology is J_1 and $c_n(x_n - e) \rightarrow y$, y has no negative jumps; e.g., see p. 301 of Jacod and Shiryaev (1987). Similarly, $c_n(x_n^{-1} - e)$ has no negative jumps. Since $c_n(x_n^{-1} - e) \rightarrow -y$ (J_1), $-y$ has no negative jumps. ■

The following lemma establishes a necessary condition in any of the topologies.

Lemma 13.7.2. *If $x_n \in D_{u,\uparrow}$, $c_n(x_n - e)(0) \rightarrow y(0)$ and $c_n(x_n^{-1} - e)(0) \rightarrow -y(0)$, where $c_n \rightarrow \infty$, then $y(0) = 0$.*

Proof. Since $x_n \in D_{u,\uparrow}$, $x_n(0) \geq 0$ and $x_n^{-1}(0) \geq 0$. Since $e(0) = 0$, the convergence $c_n(x_n - e)(0) \rightarrow y(0)$ implies that $y(0) \geq 0$. Similarly, the convergence $c_n(x_n^{-1} - e)(0) \rightarrow -y(0)$ implies that $y(0) \leq 0$. ■

Now we state the main limit theorem for inverse functions with centering.

Theorem 13.7.1. (inverse with linear centering) *Suppose that $c_n(x_n - e) \rightarrow y$ as $n \rightarrow \infty$ in $D([0, \infty), \mathbb{R})$ with one of the topologies M_2 , M_1 or J_1 , where $x_n \in D_u$, $c_n \rightarrow \infty$ and $y(0) = 0$.*

(a) *If the topology is M_2 or M_1 , then $c_n(x_n^{-1} - e) \rightarrow -y$ as $n \rightarrow \infty$ with the same topology.*

(b) *If the topology is J_1 and if y has no positive jumps, then $c_n(x_n^{-1} - e) \rightarrow -y$ as $n \rightarrow \infty$.*

We can combine Lemma 13.6.6 and Theorem 13.7.1 to obtain the following corollary. Let Λ be the space of homeomorphisms of $\mathbb{R}_+$.

Corollary 13.7.1. *Suppose that $x_n \in D_{u,\uparrow}$ and $\lambda_{1,n}, \lambda_{2,n} \in \Lambda$, $n \geq 1$. Let $c_n \rightarrow \infty$ and $y(0) = 0$. Then*

$$c_n(\lambda_{2,n} \circ x_n \circ \lambda_{1,n} - e) \rightarrow y \quad \text{in} \quad D([0, \infty), \mathbb{R}, M_i) \quad (7.1)$$

if and only if

$$c_n(\lambda_{1,n}^{-1} \circ x_n^{-1} \circ \lambda_{2,n}^{-1} - e) \rightarrow -y \quad \text{in} \quad D([0, \infty), \mathbb{R}, M_i), \quad (7.2)$$

where the topology in both cases is either M_1 or M_2 .

We can apply Corollary 13.7.1 to obtain generalizations of Theorem 13.7.1 with nonlinear centering terms. (We obtain a more general result at the end of the section.)

Corollary 13.7.2. (centering functions from Λ) *Suppose that, in addition to the conditions of Corollary 13.7.1, $\lambda_{i,n} \rightarrow \lambda_i$ as $n \rightarrow \infty$ for each i , where $\lambda_i \in \Lambda$. Then*

$$c_n(\lambda_{2,n} \circ x_n - \lambda_{1,n}^{-1}) \rightarrow y \circ \lambda_1^{-1} \quad \text{in} \quad (D, M_i) \quad (7.3)$$

if and only if

$$c_n(\lambda_{1,n}^{-1} \circ x_n^{-1} - \lambda_{2,n}^{-1}) \rightarrow -y \circ \lambda_2^{-1} \quad \text{in} \quad (D, M_i). \quad (7.4)$$

Proof. Apply Theorem 13.2.3 with the composition map to show that (7.3) is equivalent to (7.1) and (7.4) is equivalent to (7.2). ■

We can use Corollary 13.7.1 to obtain the following consequence.

Corollary 13.7.3. *Suppose that $x_n \in D_u$, $y(0) = 0$, $c_n \rightarrow \infty$ and $a_n \rightarrow a > 0$. If*

$$c_n(x_n - a_n e) \rightarrow y \quad \text{in} \quad D$$

with the M_1 or M_2 topology, then

$$c_n(x_n^{-1} - a_n^{-1} e) \rightarrow -a^{-1} y \circ a^{-1} e \quad \text{in} \quad D$$

with the same topology.

Proof. Under the condition, $(a_n c_n)(a_n^{-1} x_n - e) \rightarrow x$, so that by Corollary 13.7.1, $(a_n c_n)(x_n^{-1} \circ a_n e - e) \rightarrow -y$. Now applying the composition map with $a_n^{-1} e$, $a_n c_n(x_n^{-1} - a_n^{-1} e) \rightarrow x \circ a^{-1} e$. Dividing by a_n yields the conclusion. ■

Stochastic limit theorems are not often expressed directly in the form of Corollaries 13.7.1 or 13.7.3. We now state consequences of Corollary 13.7.1 that have more direct applications.

Corollary 13.7.4. *Let $y_n \in D_{u,\uparrow}$ and $\phi_{1,n}, \phi_{2,n} \in \Lambda$, $n \geq 1$; let $u(0) = 0$ and $n/\psi(n) \rightarrow \infty$ as $n \rightarrow \infty$. Let*

$$w_n(t) \equiv \psi(n)^{-1}[(\phi_{2,n} \circ y_n \circ \phi_{1,n})(nt) - nt], \quad t \geq 0, \quad (7.5)$$

and

$$x_n(t) \equiv \psi(n)^{-1}[(\phi_{1,n}^{-1} \circ y_n^{-1} \circ \phi_{2,n}^{-1})(nt) - nt], \quad t \geq 0, \quad (7.6)$$

for all $n \geq 1$. Then

$$w_n \rightarrow u \quad \text{in} \quad D([0, \infty), \mathbb{R}) \quad (7.7)$$

if and only if

$$x_n \rightarrow -u \quad \text{in} \quad D([0, \infty), \mathbb{R}), \quad (7.8)$$

where the topology throughout is either M_1 or M_2 .

Proof. Apply Corollary 13.7.1 with $x_n(t) = n^{-1}y_n(t)$, $\lambda_{i,n}(t) = n^{-1}\phi_{i,n}(nt)$ and $c_n = n/\psi(n)$. Then $w_n = c_n(\lambda_{2,n} \circ x_n \circ \lambda_{1,n} - e)$ and $x_n = c_n(\lambda_{1,n}^{-1} \circ x_n^{-1} \circ \lambda_{2,n}^{-1} - e)$. ■

We now consider the special case of Corollary (13.7.4) in which the homeomorphisms $\phi_{i,n}$ are linear, i.e., $\phi_{i,n} = a_{i,n}e$, $n \geq 1$.

Corollary 13.7.5. *Suppose that $y_n \in D_{u,\uparrow}$, $w(0) = 0$, $a_n \rightarrow a > 0$ and $n/\psi(n) \rightarrow \infty$ as $n \rightarrow \infty$. Let*

$$\tilde{w}_n = \psi(n)^{-1}[y_n(nt) - a_n nt], \quad t \geq 0, \quad (7.9)$$

and

$$\tilde{x}_n = \psi(n)^{-1}[y_n^{-1}(nt) - a_n^{-1} nt], \quad t \geq 0. \quad (7.10)$$

Then

$$\tilde{w}_n \rightarrow w \quad \text{in} \quad D([0, \infty), \mathbb{R}) \quad (7.11)$$

if and only if

$$\tilde{x}_n \rightarrow a^{-1}w \circ a^{-1}e \quad \text{in} \quad D([0, \infty), \mathbb{R}), \quad (7.12)$$

where the topology throughout is M_1 or M_2 .

Proof. Apply Corollary 13.7.1 with $x_n(t) = n^{-1}y_n(nt)$, $\lambda_{2,n} = a_n^{-1}e$, $\lambda_{1,n} = e$ and $c_n = na_n/\psi(n)$. Then $\tilde{w}_n = c_n(\lambda_{2,n} \circ x_n \circ \lambda_{1,n} - e)$, so that $\tilde{w}_n \rightarrow w$ if and only if $c_n(\lambda_{1,n}^{-1} \circ x_n^{-1} \circ \lambda_{2,n}^{-1} - e) \rightarrow -w$. However,

$$c_n(\lambda_{1,n}^{-1} \circ x_n^{-1} \circ \lambda_{2,n}^{-1} - e) = a_n \tilde{x}_n \circ a_n e \quad (7.13)$$

and

$$-a_n \tilde{x}_n \circ a_n e \rightarrow -w \quad \text{if and only if} \quad \tilde{x}_n \rightarrow -a^{-1}w \circ a^{-1}e. \quad \blacksquare \quad (7.14)$$

Following Puhalskii (1994), we can generalize Theorem 13.7.1 by allowing nonlinear centering terms. We present several results of this kind.

Theorem 13.7.2. *Suppose that*

$$c_n(x_n - \lambda) \rightarrow u \quad \text{as } n \rightarrow \infty \quad \text{in } D$$

with one of the topologies M_2 , M_1 or J_1 , where $x_n \in D_u$, $u(0) = 0$, u has no positive jumps if the topology is J_1 , $\lambda \in \Lambda$ and $c_n \rightarrow \infty$. Then

$$c_n(\lambda \circ x_n^{-1} - e) \rightarrow -u \circ \lambda^{-1} \quad \text{as } n \rightarrow \infty \quad (7.15)$$

with the same topology. If, in addition, λ is absolutely continuous with continuous positive derivative $\dot{\lambda}$, then

$$c_n(x_n^{-1} - \lambda^{-1}) \rightarrow \frac{-u \circ \lambda^{-1}}{\dot{\lambda} \circ \lambda^{-1}} \quad \text{as } n \rightarrow \infty, \quad (7.16)$$

where $(u/v)(t) \equiv u(t)/v(t)$, $t \geq 0$.

Proof. Apply Theorems 13.2.2 and 13.2.3 with the composition map to get $c_n(x_n \circ \lambda^{-1} - \lambda \circ \lambda^{-1}) \rightarrow u \circ \lambda^{-1}$ as in the same topology. Since $\lambda \circ \lambda^{-1} = e$, we can apply Theorem 13.7.1 or Corollary 13.7.1 to get (7.15) with the same topology. Now suppose that λ is absolutely continuous with continuous positive derivative $\dot{\lambda}$. Then

$$\begin{aligned} c_n(\lambda \circ x_n^{-1} - e)(t) &= c_n(\lambda \circ x_n^{-1} - \lambda \circ \lambda^{-1})(t) \\ &= c_n \int_{\lambda^{-1}(t)}^{x_n^{-1}(t)} \dot{\lambda}(s) ds. \end{aligned} \quad (7.17)$$

Since $c_n(x_n - \lambda) \rightarrow u$, $\|x_n - \lambda\|_t \rightarrow 0$ and $\|x_n^{-1} - \lambda^{-1}\|_t \rightarrow 0$ as $n \rightarrow \infty$ for all t . Since $\dot{\lambda}$ is continuous, it is uniformly continuous over bounded intervals. Hence

$$\sup_{0 \leq s \leq t} \left| c_n \int_{\lambda^{-1}(s)}^{x_n^{-1}(s)} \dot{\lambda}(u) du - \dot{\lambda}(\lambda^{-1}(s)) c_n(x_n^{-1}(s) - \lambda^{-1}(s)) \right| \rightarrow 0. \quad (7.18)$$

Then (7.15), (7.17) and (7.18) imply that

$$(\dot{\lambda} \circ \lambda^{-1})c_n(x_n^{-1} - \lambda^{-1}) \rightarrow -u \circ \lambda^{-1} \quad \text{as } n \rightarrow \infty \quad (7.19)$$

in the same topology, where $(uv)(t) \equiv u(t)v(t)$ for $u, v \in D$. Finally (7.19) implies (7.16). ■

Corollary 13.7.6. *Suppose that $x_n \in D_{u,\uparrow}$, $u(0) = 0$, $\lambda \in \Lambda$, λ is absolutely continuous with continuous positive derivative $\dot{\lambda}$ and $c_n \rightarrow \infty$. Then*

$$c_n(x_n - \lambda) \rightarrow u \quad \text{in } D \quad (7.20)$$

with one of the topologies M_1 or M_2 if and only if

$$c_n(x_n^{-1} - \lambda^{-1}) \rightarrow \frac{-u \circ \lambda^{-1}}{\dot{\lambda} \circ \lambda^{-1}} \quad \text{in } D \quad (7.21)$$

with the same topology.

Proof. The implication (7.20) $\rightarrow$ (7.21) is directly covered by Theorem 13.7.2. to go the other way, note that $\lambda^{-1} \in \Lambda$ and λ^{-1} is absolutely continuous with continuous positive derivative $1/\dot{\lambda}(\lambda^{-1}(t))$. Moreover, if $v = -(u \circ \lambda^{-1})/\dot{\lambda} \circ \lambda^{-1}$ in (7.21), then $v(0) = 0$ and $-(v \circ \lambda)/(\dot{\lambda}^{-1}) \circ \lambda = u$. ■

We can often apply the basic convergence-preservation results in combination. We can combine Theorems 13.3.1 and 13.7.2 to obtain limits for functions $x_n \circ y_n^{-1}$ and $x_n^{-1} \circ y_n$ with nonlinear centering.

Theorem 13.7.3. (composition plus inverse with centering) *Suppose that $x_n \in D$, $y_n \in D_u$, $c_n \rightarrow \infty$,*

$$c_n(x_n - x, y_n - y) \rightarrow (u, v) \quad \text{in } D \times D \quad (7.22)$$

with one of the J_1 , M_1 or M_2 topologies, where $v(0) = 0$ and v has no positive jumps if the topology is J_1 , $y \in \Lambda$, x and y are absolutely continuous with continuous derivative $\dot{x}$ and $\dot{y}$ with $\dot{y} > 0$ and

$$\text{Disc}(u) \cap \text{Disc}(v) = \emptyset. \quad (7.23)$$

Then

$$c_n(x_n \circ y_n^{-1} - x \circ y^{-1}) \rightarrow u \circ y^{-1} - \left(\frac{\dot{x} \circ y^{-1}}{\dot{y} \circ y^{-1}} \right) (v \circ y^{-1}) \quad \text{in } D. \quad (7.24)$$

If the topology is M_1 or M_2 , then instead of (7.23) it suffices for u and v to have no common discontinuities with jumps of common (opposite) sign with $\dot{x}(t) \geq (\leq) 0$ for all t .

Proof. The conditions imply that the conditions of Theorem 13.7.2 hold for y_n , so that

$$c_n(y_n^{-1} - y^{-1}) \rightarrow -\frac{v \circ y^{-1}}{\dot{y} \circ y^{-1}} \quad \text{in } D. \quad (7.25)$$

The conditions then imply that the conditions of Theorem 13.3.3 hold with y_n^{-1} here playing the role of y_n there. We need

$$\text{Disc}(u \circ y^{-1}) \cap \text{Disc}(v \circ y^{-1}) = \phi \quad (7.26)$$

but that is equivalent to (7.23). With the M topologies, we can apply Theorems 12.7.3 and 12.11.6 to treat addition and Theorem 13.3.2 to treat multiplication. ■

We now turn to the general first passage times

$$(x_n^{-1} \circ y_n)(t) = \inf\{s \geq 0 : x_n(s) > y_n(t)\}, \quad t \geq 0, \quad (7.27)$$

which are elements of D when $x_n \in D_u$ and $y_n \in D_\uparrow$. The following is Puhalskii's (1994) result extended to allow discontinuous limits. For an application to obtain heavy-traffic stochastic-process limits for waiting times directly from corresponding heavy-traffic stochastic-process limits for queue lengths, see Section 5.4 of the Internet Supplement.

Theorem 13.7.4. (Puhalskii's theorem) *Suppose that $x_n \in D_u$, $y_n \in D_\uparrow$, $c_n \rightarrow \infty$,*

$$c_n(x_n - x, y_n - y) \rightarrow (u, v) \quad \text{in } D \times D \quad (7.28)$$

with one of the J_1 , M_1 or M_2 topologies, where $u(0) = 0$, u has no positive jumps if the topology is J_1 ,

$$\text{Disc}(u \circ x^{-1} \circ y) \cap \text{Disc}(v) = \phi, \quad (7.29)$$

$y \in C_{\uparrow\uparrow}$ and x is absolutely continuous with a continuous positive derivative $\dot{x}$, then

$$c_n(x_n^{-1} \circ y_n - x^{-1} \circ y) \rightarrow \frac{v - u \circ x^{-1} \circ y}{\dot{x} \circ x^{-1} \circ y} \quad \text{in } D \quad (7.30)$$

with the same topology. If the topology is M_1 or M_2 , then instead of condition (7.29) it suffices for $u \circ x^{-1} \circ y$ and v to have no common discontinuities with jumps of common sign.

Proof. Since x is absolutely continuous with continuous positive derivative $\dot{x}$, $x \in C_{\uparrow\uparrow}$. Hence the conditions of Theorem 13.7.2 hold, so that

$$c_n(x_n^{-1} - x^{-1}) \rightarrow \frac{-u \circ x^{-1}}{\dot{x} \circ x^{-1}} \quad \text{in } D \quad (7.31)$$

with the same topology. We now apply Theorem 13.3.3, noting that x^{-1} has a continuous derivative $1/\dot{x}(x^{-1}(t))$. Condition (7.29) implies condition (3.14) for u in (3.14) equal to $-(u \circ x^{-1})/\dot{x} \circ x^{-1}$. Then (3.15) becomes (7.30). With the M topologies, we can apply Theorems 12.7.3 and 12.11.6. ■

Remark 13.7.1. *Relating the theorems under extra conditions.* Under extra regularity conditions, we can apply Theorem 13.7.2 to obtain limits for $y_n \circ x_n^{-1}$ from limits for $x_n \circ y_n^{-1}$ provided by Theorem 13.7.3. We need $u(0) = v(0) = 0$, $x, y \in \Lambda$, $x_n, y_n \in D_u$ and both $\dot{x}$ and $\dot{y}$ to be continuous and positive. Since $x_n, y_n \in D_u$, $x_n^{-1}, y_n^{-1} \in D_{u,\uparrow}$. Then $\lambda \equiv x \circ y^{-1} \in \Lambda$ and $(x_n \circ y_n^{-1})^{-1} = y_n \circ x_n^{-1}$. From (7.16) and (7.24), we obtain

$$c_n(y_n \circ x_n^{-1} - y \circ x^{-1}) \rightarrow z \quad (7.32)$$

where

$$z = \frac{-1}{\dot{\lambda} \circ \lambda^{-1}} \left(u \circ y^{-1} - \left(\frac{\dot{x} \circ y^{-1}}{\dot{y} \circ y^{-1}} \right) (v \circ y^{-1}) \right) \circ \lambda^{-1} \quad (7.33)$$

for $\lambda = x \circ y^{-1}$. Since $\lambda^{-1} = y \circ x^{-1}$,

$$\dot{\lambda} = \frac{\dot{x} \circ y^{-1}}{\dot{y} \circ y^{-1}}, \quad \dot{\lambda} \circ \lambda^{-1} = \frac{\dot{x} \circ x^{-1}}{\dot{y} \circ x^{-1}} \quad (7.34)$$

and

$$z = -\frac{(\dot{y} \circ x^{-1})}{\dot{x} \circ x^{-1}} (u \circ x^{-1}) + v \circ x^{-1}, \quad (7.35)$$

which coincides with (7.24) with the labels changed, i.e., with (x, y, u, v) replaced by (y, x, v, u) .

Similarly, under extra regularity conditions, we can apply Theorem 13.7.2 to obtain limits for $y_n^{-1} \circ x_n$ from limits for $x_n^{-1} \circ y_n$ provided by Theorem 13.7.4. We now need $x_n, y_n \in D_{u,\uparrow}$. We obtain

$$c_n(y_n^{-1} \circ x_n - y^{-1} \circ x) \rightarrow z, \quad (7.36)$$

where

$$z = \frac{-1}{\dot{\lambda} \circ \lambda^{-1}} \left(\frac{v - u \circ x^{-1} \circ y}{\dot{x} \circ x^{-1} \circ y} \right) \circ \lambda^{-1} \quad (7.37)$$

for $\lambda = x^{-1} \circ y$. Since $\lambda^{-1} = y^{-1} \circ x$,

$$\dot{\lambda} = \frac{\dot{y}}{\dot{x} \circ x^{-1} \circ y}, \quad \dot{\lambda} \circ \lambda^{-1} = \frac{\dot{y} \circ y^{-1} \circ x}{\dot{x}}, \quad (7.38)$$

and

$$z = \frac{u - v \circ y^{-1} \circ x}{\dot{y} \circ y^{-1} \circ x}, \quad (7.39)$$

which agrees with (7.30) with the labels changed, i.e., with (x, y, u, v) replaced by (y, x, v, u) . ■

13.8. Counting Functions

Inverse functions or first-passage-time functions are closely related to counting functions. A counting function is defined in terms of a sequence $\{s_n : n \geq 0\}$ of nondecreasing nonnegative real numbers with $s_0 = 0$. We can think of s_n as the partial sum

$$s_n \equiv x_1 + \cdots + x_n, \quad n \geq 1, \quad (8.1)$$

by simply writing $x_i \equiv s_i - s_{i-1}$, $i \geq 1$. The associated *counting function* $\{c(t) : t \geq 0\}$ is defined by

$$c(t) \equiv \max\{k \geq 0 : s_k \leq t\}, \quad t \geq 0. \quad (8.2)$$

To have $c(t)$ finite for all $t > 0$, we assume that $s_n \rightarrow \infty$ as $n \rightarrow \infty$. We can reconstruct the sequence $\{s_n\}$ from $\{c(t) : t \geq 0\}$ by

$$s_n = \inf\{t \geq 0 : c(t) \geq n\}, \quad n \geq 0. \quad (8.3)$$

The sequence $\{s_n\}$ and the associated function $\{c(t) : t \geq 0\}$ can serve as sample paths for a stochastic point process on the nonnegative real line. Then there are (countably) infinitely many points with the n^{th} point being located at s_n . The summands x_n are then the intervals between successive points. The most familiar case is when the sequence $\{x_n : n \geq 1\}$ constitutes the possible values from a sequence $\{X_n : n \geq 1\}$ of IID random variables with values in $\mathbb{R}_+$. Then the counting function $\{c(t) : t \geq 0\}$ constitutes a possible sample path of an associated renewal counting process $\{C(t) : t \geq 0\}$; see Section 7.3.

Paralleling Lemma 13.6.3, we have the following basic inverse relation for counting functions.

Lemma 13.8.1. (inverse relation) *For any nonnegative integer n and non-negative real number t ,*

$$s_n \leq t \quad \text{if and only if} \quad c(t) \geq n. \quad (8.4)$$

We can put counting functions in the setting of inverse functions on $D_\uparrow$ by letting

$$y(t) \equiv s_{[t]}, t \geq 0. \quad (8.5)$$

To have $y \in D_\uparrow$, we use the assumption that $s_n \rightarrow \infty$ as $n \rightarrow \infty$. if all the summands are strictly positive, then

$$y^{-1}(t) = c(t) + 1, \quad t \geq 0, \quad (8.6)$$

where y^{-1} is the image of the inverse map in (6.1) applied to y in (8.5). With (8.6), limits for counting functions can be obtained by applying results in the previous two sections.

The connection to the inverse map can also be made when the summands x_i are only nonnegative. To do so, we observe that the counting function c is a time-transformation of y^{-1} . both are right-continuous, but $c(t) < y^{-1}(t)$. In particular, c and y can be expressed in terms of each other.

Lemma 13.8.2. (relation between counting functions and inverse functions) *For y in (8.5) and c in (8.2),*

$$c(t) = y^{-1}(y(y^{-1}(t)-)-), \quad t \geq 0, \quad (8.7)$$

$$c(t) = y^{-1}(t-) \quad \text{for all } t \in \text{Disc}(c) = \text{Disc}(y^{-1}), \quad (8.8)$$

$$y^{-1}(t) = c(c^{-1}(c(t))), \quad t \geq 0. \quad (8.9)$$

The three functions y , y^{-1} and c are depicted for a typical initial segment of a sequence $\{s_n : n \geq 0\}$ in Figure 13.1.

We can apply (8.7)–(8.9) in Lemma 13.8.1 to show that limits for scaled counting functions with centering, are equivalent to limits for scaled inverse functions. We use the fact that the M topologies are not altered by changing to the left limits, because the graph is unchanged. We first consider the case of no centering; afterwards we consider the case of centering. When there is no centering, the M_1 and M_2 topologies coincide and reduce to pointwise convergence on a dense subset of $\mathbb{R}_+$ including 0.

Consider a sequence of counting functions $\{c_n(t) : t \geq 0\} : n \geq 1\}$ with associated processes

$$y_n^{-1}(t) \equiv c_n(c_n^{-1}(c_n(t))), \quad t \geq 0, \quad (8.10)$$

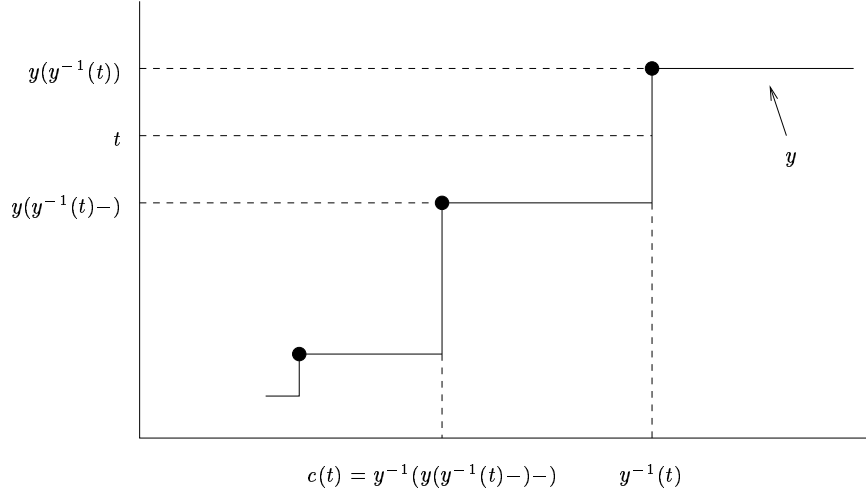


Figure 13.1: The relation between the counting function c and the inverse function y^{-1} for a typical function y .

$y_n = (y_n^{-1})^{-1}$. Form scaled functions by setting

$$\hat{c}_n(t) = n^{-1}c_n(a_nt) \quad \text{and} \quad \hat{y}_n(t) = a_n^{-1}y_n(nt), \quad t \geq 0, \quad (8.11)$$

where a_n are positive real numbers with $a_n \rightarrow \infty$. Note that

$$\hat{c}_n^{-1}(t) = a_n^{-1}c_n^{-1}(nt) \quad \text{and} \quad \hat{y}_n^{-1}(t) = n^{-1}y_n(a_nt), \quad t \geq 0. \quad (8.12)$$

Theorem 13.8.1. (asymptotic equivalence of limits for scaled processes)
Suppose that $\hat{y}_n \in D_{u,\uparrow}$, $n \geq 1$, for $\hat{y}_n$ in (8.11). Then any one of the limits $\hat{y}_n \rightarrow \hat{y}$, $\hat{y}_n^{-1} \rightarrow \hat{y}^{-1}$, $\hat{c}_n \rightarrow \hat{y}^{-1}$ or $\hat{c}_n^{-1} \rightarrow \hat{y}$ in $D_{\uparrow}([0, \infty), \mathbb{R})$ with the M_2 ($= M_1$) topology, for $\hat{y}_n^{-1}$, $\hat{c}_n$ and $\hat{c}_n^{-1}$ in (8.11) and (8.12), implies the others.

We now apply the results for inverse maps with centering in Section 13.7 to obtain limits for counting functions with centering. Consider a sequence of counting functions $\{c_n(t) : t \geq 0\} : n \geq 1\}$ associated with a sequence of nondecreasing sequences of nonnegative numbers $\{s_{n,k} : k \geq 0\} : n \geq 1\}$ defined as in (8.2). Let the scaled functions $\hat{c}_n$, $\hat{y}_n$, $\hat{c}_n^{-1}$ and $\hat{y}_n^{-1}$ be defined as in (8.10)–(8.12).

Theorem 13.8.2. (asymptotic equivalence of counting and inverse functions with centering) *Consider $\hat{\mathbf{y}}_n$, $\hat{\mathbf{c}}_n$, and $\hat{\mathbf{y}}_n^{-1}$ and $\hat{\mathbf{c}}_n^{-1}$ as defined in (8.11) and (8.12). Suppose that $\hat{\mathbf{y}}_n \in D_{u,\uparrow}$, $n \geq 1$, $b_n \rightarrow \infty$ and $\mathbf{z}(0) = 0$. Then any one of the limits $b_n(\hat{\mathbf{y}}_n - e) \rightarrow \mathbf{z}$, $b_n(\hat{\mathbf{c}}_n - e) \rightarrow -\mathbf{z}$, $b_n(\hat{\mathbf{y}}_n^{-1} - e) \rightarrow -\mathbf{z}$ or $b_n(\hat{\mathbf{c}}_n^{-1} - e) \rightarrow \mathbf{z}$ in $D([0, \infty), \mathbb{R})$ with the M_1 or M_2 topology implies the others with the same topology.*

Corollary 13.8.1. *Consider a sequence of nondecreasing nonnegative sequences $\{\{s_{n,k} : k \geq 0\} : n \geq 1\}$ with $s_{n,0} = 0$ and $s_{n,k} \rightarrow \infty$ as $k \rightarrow \infty$ for all n . Let*

$$\mathbf{x}_n(t) = \delta_n^{-1}[s_{n,\lfloor nt \rfloor} - m_n nt], \quad t \geq 0,$$

and

$$\mathbf{y}_n(t) = \delta_n^{-1}[c_n(nt) - m_n^{-1}nt], \quad t \geq 0,$$

for $c_n(t)$ defined as in (8.2). Suppose that $\mathbf{u}(0) = 0$, $\delta_n \rightarrow \infty$, $n/\delta_n \rightarrow \infty$ and $m_n \rightarrow m > 0$ as $n \rightarrow \infty$. Then $\mathbf{x}_n \rightarrow \mathbf{u}$ in $D([0, \infty), \mathbb{R})$ with the M_1 or M_2 topology if and only if $\mathbf{y}_n \rightarrow -m^{-1}\mathbf{u} \circ m^{-1}\mathbf{e}$ in $D([0, \infty), \mathbb{R})$ with the same topology.

Proof. Apply Theorem 13.8.2, letting $\hat{\mathbf{c}}_n(t) = (a_n n)^{-1}c_n(nt)$ for $a_n = m_n^{-1}$ and, necessarily, $\hat{\mathbf{y}}_n(t) = n^{-1}s_{n,\lfloor a_n nt \rfloor}$. Then $b_n(\hat{\mathbf{y}}_n - e) \rightarrow \mathbf{z}$ if and only if $b_n(\hat{\mathbf{c}}_n - e) \rightarrow -\mathbf{z}$ for $b_n \rightarrow \infty$ and $\mathbf{z}(0) = 0$. However, $b_n(\hat{\mathbf{y}}_n - e) \rightarrow \mathbf{u} \circ m^{-1}\mathbf{e}$ if and only if $\mathbf{x}_n \rightarrow \mathbf{u}$, while $b_n(\hat{\mathbf{c}}_n - e) \rightarrow m^{-1}\mathbf{z}$ if and only if $\mathbf{y}_n \rightarrow \mathbf{z}$, for $b_n = n/\delta_n \rightarrow \infty$. ■

Internet Supplement

Internet Supplement (last updated on June 21, 2001) to the book,

[Stochastic-Process Limits](#),

An Introduction to Stochastic-Process Limits
And their Application to Queues

published by Springer in 2002

- The Full Internet Supplement (about 325 pages) [\[Postscript\]](#) [\[PDF\]](#)

Drafts of individual chapters in the Internet Supplement:

- Cover, Preface and Contents [\[Postscript\]](#) [\[PDF\]](#)
- Chapter 1: Fundamentals [\[Postscript\]](#) [\[PDF\]](#)
- Chapter 2: Stochastic-Process Limits [\[Postscript\]](#) [\[PDF\]](#)
- Chapter 3: Preservation of Pointwise Convergence [\[Postscript\]](#) [\[PDF\]](#)
- Chapter 4: An Application to Simulation [\[Postscript\]](#) [\[PDF\]](#)
- Chapter 5: Heavy-Traffic Limits for Queues [\[Postscript\]](#) [\[PDF\]](#)
- Chapter 6: The Space D [\[Postscript\]](#) [\[PDF\]](#)
- Chapter 7: Useful Functions [\[Postscript\]](#) [\[PDF\]](#)
- Chapter 8: Queueing Networks [\[Postscript\]](#) [\[PDF\]](#)
- Chapter 9: Nonlinear Centering and Derivatives [\[Postscript\]](#) [\[PDF\]](#)
- Chapter 10: More on the Spaces E and F [\[Postscript\]](#) [\[PDF\]](#)
- Chapter 11: Errors Discovered in the Book [\[Postscript\]](#) [\[PDF\]](#)
- Bibliography [\[Postscript\]](#) [\[PDF\]](#)
- Index [\[Postscript\]](#) [\[PDF\]](#)

Internet Supplement to Stochastic-Process Limits

**An Introduction to Stochastic-Process
And their Application to Queues**

Ward Whitt

AT&T Labs - Research
The Shannon Laboratory
Florham Park, New Jersey

Draft
June 21, 2001 Copyright ©info

Preface

0.1. Why is there an Internet Supplement?

This Internet Supplement has three purposes: First, it is intended to maintain a list of corrections for errors found after the book has been published. Second, it is intended to provide supporting details, such as proofs, for material in the book. Third, it is intended to provide supplementary material related to the subject of the book.

As indicated in the Preface to the book, in order to avoid excessive length, material was deleted from the book and placed in this Internet Supplement. The first choice for cutting was the more technical material. Thus, the Internet Supplement contains many proofs for theorems in the book. Specifically, missing proofs for results stated in the book are contained here in Chapter 1 (all but Section 1.4), Section 5.3 and Chapters 6–8.

It was also considered necessary to cut some entire discussions. Hence the book also contains supplementary material related to, but going beyond, what is in the book. Such material is contained here in Section 1.4, Chapters 2–5 (all but Section 5.3) and Chapter 9.

In addition to making corrections as errors are discovered, the Internet Supplement provides an opportunity to add other material after the book has been published. We would like to add additional material on the spaces E and F , going beyond the brief introduction in Chapter 15 of the book.

0.2. Organization

We now indicate how the Internet Supplement is organized.

Chapter 1 here complements Chapter 3 of the book on the framework for stochastic-process limits. Sections 1.2 and 1.3 provide proofs for the Prohorov metric properties and the Skorohod representation theorem from

Section 3.2 of the book. Section 1.4 explains the adjective “weak” in “weak convergence” from a Banach-space perspective. Finally, Section 1.5 gives proofs of the continuous-mapping theorems and the Lipschitz-mapping theorem in Section 3.4 of the book.

Chapter 2 here complements Chapter 4 of the book on basic stochastic-process limits. Section 2.2 complements Section 4.3 of the book on Donsker’s theorem by providing an introduction to strong approximations and their application to establish rates of convergence in the setting of Donsker’s theorem, using the Prohorov metric on the space of probability measures $\mathcal{P}$ on the function space D . Section 2.3 complements Section 4.4 of the book on Brownian limits with weak dependence by presenting FCLT’s exploiting Markov, regenerative and martingale structure. Section 2.4 complements Section 4.5 in the book on convergence to stable Lévy motion by discussing FCLT’s in the framework of double sequences (or triangular arrays) of random variables; with an IID assumption, the scaled partial sums converge to general Lévy processes. Finally, Section 2.5 complements Section 4.6 of the book on strong dependence by showing that the linear-process representation in equation (6.6) of the book arises naturally in the framework of time-series models.

Chapter 3 here complements Chapter 13 of the book on useful functions that preserve convergence by showing how pointwise convergence in $\mathbb{R}$ is preserved under mappings. Section 3.2 shows that in some settings pointwise convergence directly implies uniform convergence over bounded intervals. As a consequence, an ordinary strong law of large numbers (SLLN) directly implies the more general functional strong law of large numbers (FSLLN). The remaining sections in Chapter 3 discuss the preservation of pointwise convergence under the supremum, inverse and composition maps. With the inverse map, attention is focused on counting processes, with and without centering.

Chapter 4 here complements Sections 5.9 and 10.4.4 of the book by discussing another application of stochastic-process limits to simulation. Sections 5.9 and 10.4.4 of the book show how heavy-traffic stochastic-process limits for queues can be used to help plan queueing simulations. In particular, they determine the approximate required simulation run length, as a function of model parameters, in order to achieve desired statistical precision. Drawing upon and extending Glynn and Whitt (1992a), Chapter 4 shows how FCLT’s and the continuous-mapping approach can be used to establish general criteria for sequential stopping rules for simulations to be asymptotically valid.

Chapter 5 here complements Chapters 5, 8 and 9 of the book on single-

server queues. Section 5.2 here discusses general reflected-Lévy-process approximations for queues that arise when there is a sequence of queueing models with net-input processes satisfying the FCLT's discussed here in Section 2.4. Section 5.3 here provides the proof of Theorem 8.3.1 in the book, which establishes a FCLT for the cumulative busy time of a single on-off source. Finally, following Puhalskii (1994), Section 5.4 here shows how the continuous-mapping approach with the inverse map and nonlinear centering in Theorem 13.7.4 of the book can be used to convert stochastic-process limits for arrival, departure and queue-length processes into associated stochastic-process limits for waiting-time and workload processes in quite general queueing models.

Chapters 6, 7 and 8 here provide proofs for theorems in Chapters 12, 13 and 14, respectively, in the book. The numbering within the chapters here closely parallels the numbering within the corresponding chapter in the book, so the desired proof here should be easy to find. In addition, there is an extra section in Chapter 8 here on queueing networks. Drawing on and extending Kella and Whitt (1996), Section 8.9 establishes general conditions for a multidimensional reflected process to have a limiting stationary version.

Chapter 9 here continues the study of useful functions begun in Chapter 13 of the book. In particular, drawing upon and extending Mandelbaum and Massey (1995), Chapter 9 here studies convergence preservation of the supremum, (one-sided, one-dimensional) reflection and inverse maps with nonlinear centering. Under regularity conditions, the limit for the scaled functions after applying these maps can be identified with an appropriate “directional” derivative of the map.

Finally, Chapter 10 here is intended to contain corrections for errors found after the book has been published.

Contents

Preface	iii
0.1 Why is there an Internet Supplement?	iii
0.2 Organization	iii
1 Fundamentals	1
1.1 Introduction	1
1.2 The Prohorov Metric	1
1.3 The Skorohod Representation Theorem	6
1.3.1 Proof for the Real Line	6
1.3.2 Proof for Complete Separable Metric Sspaces	7
1.3.3 Proof for Separable Metric Spaces	10
1.4 The “Weak” in Weak Convergence	16
1.5 Continuous-Mapping Theorems	17
1.5.1 Proof of the Lipschitz Mapping Theorem	17
1.5.2 Proof of the Continuous-Mapping Theorems	19
2 Stochastic-Process Limits	23
2.1 Introduction	23
2.2 Strong Approximations and Rates of Convergence	23
2.2.1 Rates of Convergence in the CLT	24
2.2.2 Rates of Convergence in the FCLT	25
2.2.3 Strong Approximations	27
2.3 Weak Dependence from Regenerative Structure	30
2.3.1 Discrete-Time Markov Chains	31
2.3.2 Continuous-Time Markov Chains	33
2.3.3 Regenerative FCLT	36
2.3.4 Martingale FCLT	40
2.4 Double Sequences and Lévy Limits	41
2.5 Linear Models	46

3	Preservation of Pointwise Convergence	51
3.1	Introduction	51
3.2	From Pointwise to Uniform Convergence	52
3.3	Supremum	54
3.4	Counting Functions	55
3.5	Counting Functions with Centering	62
3.6	Composition	68
3.7	Chapter Notes	71
4	An Application to Simulation	73
4.1	Introduction	73
4.2	Sequential Stopping Rules for Simulations	73
4.2.1	The Mathematical Framework	75
4.2.2	The Absolute-Precision Sequential Estimator	79
4.2.3	The Relative-Precision Sequential Estimator	82
4.2.4	Analogs Based on a FWLLN	83
4.2.5	Examples	86
5	Heavy-Traffic Limits for Queues	97
5.1	Introduction	97
5.2	General Lévy Approximations	97
5.3	A Fluid Queue Fed by On-Off Sources	100
5.3.1	Two False Starts	101
5.3.2	The Proof	103
5.4	From Queue Lengths to Waiting Times	105
5.4.1	The Setting	105
5.4.2	The Inverse Map with Nonlinear Centering	106
5.4.3	An Application to Central-Server Models	110
6	The Space D	113
6.1	Introduction	113
6.2	Regularity Properties of D	114
6.3	Strong and Weak M_1 Topologies	117
6.3.1	Definitions	117
6.3.2	Metric Properties	118
6.3.3	Properties of Parametric Representations	122
6.4	Local Uniform Convergence at Continuity Points	124
6.5	Alternative Characterizations of M_1 Convergence	128
6.5.1	SM_1 Convergence	128
6.5.2	WM_1 Convergence	130

6.6	Strengthening the Mode of Convergence	134
6.7	Characterizing Convergence with Mappings	134
6.7.1	Linear Functions of the Coordinates	135
6.7.2	Visits to Strips	137
6.8	Topological Completeness	139
6.9	Non-Compact Domains	142
6.10	Strong and Weak M_2 Topologies	144
6.10.1	The Hausdorff Metric Induces the SM_2 Topology . . .	145
6.10.2	WM_2 is the Product Topology	147
6.11	Alternative Characterizations of M_2 Convergence	148
6.11.1	M_2 Parametric Representations	148
6.11.2	SM_2 Convergence	148
6.11.3	WM_2 Convergence	155
6.11.4	Additional Properties of M_2	158
6.12	Compactness	161
7	Useful Functions	163
7.1	Introduction	163
7.2	Composition	164
7.2.1	Preliminary Results	164
7.2.2	M -Topology Results	166
7.3	Composition with Centering	173
7.4	Supremum	173
7.4.1	The Supremum without Centering	173
7.4.2	The Supremum with Centering	174
7.5	One-Dimensional Reflection	181
7.6	Inverse	183
7.6.1	The M_1 Topology	183
7.6.2	The M'_1 Topology	186
7.7	Inverse with Centering	188
7.8	Counting Functions	190
7.9	Renewal-Reward Processes	194
8	Queueing Networks	195
8.1	Introduction	195
8.2	The Multidimensional Reflection Map	197
8.2.1	Definition and Characterization	197
8.2.2	Continuity and Lipschitz Properties	200
8.3	The Instantaneous Reflection Map	204
8.4	Reflections of Parametric Representations	204

8.5	M_1 Continuity Results	210
8.6	Limits for Stochastic Fluid Networks	217
8.7	Queueing Networks with Service Interruptions	217
8.8	The Two-Sided Regulator	217
8.9	Existence of a Limiting Stationary Version	218
8.9.1	The Main Results	218
8.9.2	Proofs	224
9	Nonlinear Centering and Derivatives	235
9.1	Introduction	235
9.2	Nonlinear Centering and Derivatives	237
9.3	Derivative of the Supremum Function	243
9.4	Extending Pointwise Convergence to M_1 Convergence	254
9.5	Derivative of the Reflection Map	258
9.6	Heavy-Traffic Limits for Nonstationary Queues	262
9.7	Derivative of the Inverse Map	267
9.8	Chapter Notes	276
10	Errors Discovered in the Book	281
11	Bibliography	283

Chapter 1

Fundamentals

1.1. Introduction

In this chapter we present material supplementing the book on fundamental topics. In Sections 1.2 and 1.3 we give detailed proofs of the Prohorov metric properties and the Skorohod representation theorem, stated in Theorems 3.2.1 and 3.2.2 of the book. In Section 1.4 we explain the adjective “weak” in weak convergence from a Banach-space perspective. In Section 1.5 we provide proofs of the continuous mapping theorems, stated in Section 3.4 of the book.

1.2. The Prohorov Metric

In this section we prove Theorem 3.2.1 in the book, establishing that the Prohorov (1956) metric is indeed a metric inducing weak convergence $P_n \Rightarrow P$.

Recall that we are considering probability measures on a separable metric space (S, m) . In that setting, $P_n \Rightarrow P$ if

$$\lim_{n \rightarrow \infty} \int_S f dP_n = \int_S f dP \quad (2.1)$$

for all functions f in $C(S)$, the space of all continuous bounded real-valued functions on S . Recall that the Prohorov metric π is defined on the space $\mathcal{P} \equiv \mathcal{P}(S)$ of all probability measures on the separable metric space (S, m) by

$$\pi(P_1, P_2) \equiv \inf\{\epsilon > 0 : P_1(A) \leq P_2(A^\epsilon) + \epsilon \text{ for all } A \in \mathcal{B}(S)\} \text{ , } \quad (2.2)$$

for $P_1, P_2 \in \mathcal{P}(S)$, where A^ϵ is the *open ϵ -neighborhood* of A , i.e.,

$$A^\epsilon \equiv \{y \in S : m(x, y) < \epsilon \text{ for some } x \in A\} . \quad (2.3)$$

Here is the result that we wish to prove:

Theorem 1.2.1. (the Prohorov metric on $\mathcal{P}$) *For any separable metric space (S, m) , the function π on $\mathcal{P}(S)$ in (2.2) is a separable metric. There is convergence $\pi(P_n, P) \rightarrow 0$ in $\mathcal{P}(S)$ if and only if $P_n \Rightarrow P$, as defined in (2.1). Moreover, in (2.2) it suffices to let the sets A be closed.*

To carry out the proof, we show that weak convergence $P_n \Rightarrow P$ implies uniform convergence of integrals $\int g dP_n$ for an appropriate class of functions g .

Consider a class $\mathcal{G}$ real-valued functions on S . We say that $\mathcal{G}$ is *uniformly bounded* if

$$\sup_{g \in \mathcal{G}, x \in S} \{|g(x)|\} < \infty .$$

We say that $\mathcal{G}$ is *equicontinuous at x* if, for all $\epsilon > 0$, there is a $\delta > 0$ such that

$$\sup_{g \in \mathcal{G}} |g(x) - g(y)| < \epsilon \text{ when } d(x, y) < \delta .$$

We say that $\mathcal{G}$ is *equicontinuous* if it is equicontinuous at all $x \in S$.

Lemma 1.1. (uniform convergence for a class of integrals) *Suppose that $P_n \Rightarrow P$ on a separable metric space (S, m) . Let $\mathcal{G}$ be a uniformly bounded class of measurable real-valued functions on S that is equicontinuous at all $x \in E^c$. If $P(E) = 0$, then*

$$\sup_{g \in \mathcal{G}} \left| \int g dP_n - \int g dP \right| \rightarrow 0 \text{ as } n \rightarrow \infty . \quad (2.4)$$

Proof. If (2.4) were to fail, then there must exist $\epsilon > 0$ and a sequence $\{g_n : n \geq 1\}$ of functions in $\mathcal{G}$ for which $|\int g_n dP_n - \int g_n dP| > \epsilon$ infinitely often. We will show that cannot happen. Given $P_n \Rightarrow P$, we can apply the Skorohod representation theorem to construct S -valued random elements X_n and X with probability laws P_n and P such that $X_n \rightarrow X$ w.p.1. By the almost-sure equicontinuity of $\mathcal{G}$ with respect to P ,

$$\sup_n |g_n(X_n) - g_n(X)| \rightarrow 0 \text{ w.p.1} .$$

By the uniform-boundedness condition and the bounded convergence theorem,

$$\sup_n |Eg_n(X_n) - Eg_n(X)| \leq E \left[\sup_n |g_n(X_n) - g_n(X)| \right] \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

or, equivalently,

$$\sup_n \left| \int g_n dP_n - \int g_n dP \right| \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Since that is a contradiction, (2.4) must actually hold. ■

We now define a generalization of the Prohorov metric on the space $\mathcal{P}(S)$ of all probability measures on (S, m) . We define a family of metrics indexed by the scalar γ ; the standard Prohorov metric is the special case with $\gamma = 1$. For any $P_1, P_2 \in \mathcal{P}(S)$ and $\gamma > 0$, let

$$\pi_\gamma(P_1, P_2) \equiv \inf\{\epsilon > 0 : P_1(F) \leq P_2(F^\epsilon) + \gamma\epsilon \quad \text{for all closed } F \text{ in } S\}, \quad (2.5)$$

where F^ϵ is the open ϵ -neighborhood of F , as in (2.3).

Here is our main result.

Theorem 1.1. (generalized Prohorov metric) *Let (S, m) be a separable metric space. For each $\gamma > 0$, $(\mathcal{P}(S), \pi_\gamma)$ for π_γ in (2.5) is a separable metric space. The definition is unchanged if the closed sets F in (2.5) are replaced by general measurable sets A . There is convergence $\pi_\gamma(P_n, P) \rightarrow 0$ as $n \rightarrow \infty$ if and only if $P_n \Rightarrow P$.*

In preparation for the proof, we first establish some preliminary results. We first show that $\pi_\gamma(P_2, P_1) = \pi_\gamma(P_1, P_2)$. For that purpose, use the following elementary lemma. Recall that A^- is the closure of the set A .

Lemma 1.2. *For any subset A of S and $\alpha > 0$,*

$$A^- = S - (S - A^\alpha)^\alpha. \quad (2.6)$$

Lemma 1.3. *If $P_1(F) \leq P_2(F^\alpha) + \beta$ for all closed F for $\alpha, \beta > 0$, then $P_2(F) \leq P_1(F^\alpha) + \beta$ for all closed F .*

Proof. Since F^α is open, $S - F^\alpha$ is closed. Under the condition,

$$P_1(S - F^\alpha) \leq P_2((S - F^\alpha)^\alpha) + \beta ,$$

so that

$$P_2(S - (S - F^\alpha)^\alpha) \leq P_1(F^\alpha) + \beta .$$

By Lemma 1.2, $F = S - (S - F^\alpha)^\alpha$. hence

$$P_2(F) = P_2(S - (S - F^\alpha)^\alpha) \leq P_1(F^\alpha) + \beta . \quad \blacksquare$$

We now show that closed sets and measurable sets are interchangeable in (2.5).

Lemma 1.4. (closed sets suffice) *For any constants $\alpha > 0$ and $\beta > 0$, the inequality $P_1(A) \leq P_2(A^\alpha) + \beta$ holds for all $A \in \mathcal{S}$ if and only if it holds for all $A = F$, where F is closed.*

Proof. One direction is immediate. For the non-trivial direction, given any measurable set A , choose a sequence of closed sets $\{F_n : n \geq 1\}$ such that $F_n \subseteq F_{n+1}$ and $F_n \uparrow A$. Then $F_n^\alpha \uparrow F^\alpha$, $P_1(F_n) \uparrow P_1(A)$ and $P_2(F_n^\alpha) \uparrow P_2(A^\alpha)$. Hence we have $P_1(A) \leq P_2(A^\alpha) + \beta$ when we have $P_1(F_n) \leq P_2(F_n^\alpha) + \beta$ for all n . $\blacksquare$

Proof of Theorem 1.1. Lemma 1.3 establishes the symmetry property. if $\pi_\gamma(P_1, P_2) = 0$, then $P_1(F) = P_2(F)$ for each closed subset F . Since the closed sets form a determining class, $P_1 = P_2$. To establish the triangle inequality, suppose that $\pi_\gamma(P_1, P_2) < \epsilon_1 < \pi_\gamma(P_1, P_2) + \delta$ and $\pi_\gamma(P_2, P_3) < \epsilon_2 < \pi_\gamma(P_2, P_3) + \delta$ for some $\delta > 0$. Then for any closed F ,

$$\begin{aligned} P_1(F) &\leq P_2(F^{\epsilon_1}) + \gamma\epsilon_1 \\ &\leq P_2((F^{\epsilon_1})^-) + \gamma\epsilon_1 \\ &\leq P_3(F^{\epsilon_1+\epsilon_2}) + \gamma(\epsilon_1 + \epsilon_2) , \end{aligned}$$

so that

$$\pi_\gamma(P_1, P_3) \leq \epsilon_1 + \epsilon_2 \leq \pi_\gamma(P_1, P_2) + \pi_\gamma(P_2, P_3) + 2\delta .$$

Since δ was arbitrary, the triangle inequality is established, completing the proof of the metric property.

If $\pi_\gamma(P_n, P) \rightarrow 0$, then for any $\epsilon > 0$ there exists n_0 such that $P_n(F) \leq P(F^\epsilon) + \gamma\epsilon$ for all closed F and $n \geq n_0$. Hence

$$\limsup_{n \rightarrow \infty} P_n(F) \leq P(F^\epsilon) + \gamma\epsilon .$$

However, $F^\epsilon \downarrow F$ as $\epsilon \downarrow 0$, so that $P(F^\epsilon) \downarrow P(F)$ as $\epsilon \downarrow 0$. Hence,

$$\limsup_{n \rightarrow \infty} P_n(F) \leq P(F) ,$$

which implies $P_n \Rightarrow P$ by Theorem 11.3.1 in the book.

Next we show that $\pi_\gamma(P_n, P) \rightarrow 0$ if $P_n \Rightarrow P$. For each $A \in \mathcal{S}$, define

$$g_A(x) \equiv [1 - \epsilon^{-1}m(x, A)]^+ . \quad (2.7)$$

Notice that $I_A(x) \leq g_A(x) \leq I_{A^\epsilon}(x)$ for all x , where I_B is the indicator function of the set B . Moreover,

$$|g_A(x) - g_A(y)| \leq \epsilon^{-1}|m(x, A) - m(y, A)| \leq \epsilon^{-1}m(x, y)$$

for all A , so that the class of all such g_A defined in (2.7) is uniformly bounded and equicontinuous. By Lemma 1.1,

$$\Delta_n \equiv \sup_{A \in \mathcal{S}} \left| \int g_A dP_n - \int g_A dP \right| \rightarrow 0 \quad \text{as } n \rightarrow \infty .$$

Then

$$P(A^\epsilon) \geq \int g_A dP \geq \int g_A dP_n - \Delta_n \geq P_n(A) - \Delta_n$$

so that

$$P_n(A) \leq P(A^\epsilon) + \epsilon$$

when $\Delta_n < \epsilon$.

Finally, we want to show that $(\mathcal{P}(S), \pi_\gamma)$ is separable. For that purpose, let S_0 be a countable dense subset of (S, m) , which exists because we have assumed that (S, m) is separable. We will show that the countable family of rational-valued probability measures with finite support in S_0 are dense in $\mathcal{P}(S)$.

Given any $P_1 \in \mathcal{P}(S)$ and any $\epsilon > 0$, we show how to construct P_2 with finite support in S_0 such that $P_1(A) \leq P_2(A^\epsilon)$ for all $A \in \mathcal{S}$, so that $\pi_\gamma(P_1, P_2) \leq \epsilon$. Let the sequence $\{x_n : n \geq 1\}$ enumerate the elements of S_0 . We construct a partition of S containing subset of ϵ -balls about points in S_0 . We start by letting $C_1 = B_m(x_1, \epsilon)$. For $C_1, \dots, C_n$ given, let k_{n+1} be the index of the first point from $\{x_n : n \geq 0\}$ not contained in $\cup_{i=1}^n C_i$. Then let

$$C_{n+1} = B_m(x_{k_{n+1}}, \epsilon) - \cup_{i=1}^n C_i .$$

Let $k_1 = 1$. Now let P_2 attach mass $P_1(C_n)$ to point x_{k_n} (in C_n) for $n \geq 1$. To give P_2 finite support, stop when $P_1(\cup_{i=1}^k C_i) > 1 - \gamma\epsilon$ and let P_2

assign the mass $P_1(\cup_{k+1}^{\infty} C_i)$ to x_1 . Hence $P_2(\{x_1\}) = P_1(C_1) + P_1(\cup_{k+1}^{\infty} C_i)$. Now consider an arbitrary measurable set A . Note that $C_i \subseteq A^c$ whenever $A \cap C_i \neq \phi$. Since $\{C_i\}$ is a partition of S ,

$$P_1(A) = \sum_{i=1}^{\infty} P_1(A \cap C_i) \leq \sum_{i=1}^k P_1(A \cap C_i) + \gamma\epsilon \leq P_2(A^c) + \gamma\epsilon. \quad \blacksquare$$

1.3. The Skorohod Representation Theorem

In this section we prove the Skorohod representation theorem, Theorem 3.2.2 in the book. We restate it here:

Theorem 1.3.1. (Skorohod representation theorem) *If $X_n \Rightarrow X$ in a separable metric space (S, m) , then there exist other random elements of (S, m) , $\tilde{X}_n, n \geq 1$, and $\tilde{X}$, defined on a common underlying probability space, such that*

$$\tilde{X}_n \stackrel{d}{=} X_n, n \geq 1, \quad \tilde{X} \stackrel{d}{=} X$$

and

$$P(\lim_{n \rightarrow \infty} \tilde{X}_n = \tilde{X}) = 1.$$

We start by giving an elementary proof for the case in which the space S is the real line. Then we give Skorohod's (1956) original proof for the case in which S is a complete separable metric space. Finally, we give a proof for general separable metric spaces due to Wichura (1970). Dudley (1968) first showed that the completeness condition is not needed.

1.3.1. Proof for the Real Line

Suppose that $S = \mathbb{R}$. Then we can characterize the probability laws of X and $X_n, n \geq 1$, by their cumulative distribution functions (cdf's), i.e.,

$$F(t) \equiv P(X \leq t), \quad t \in \mathbb{R}. \quad (3.1)$$

For any cdf F , let F be its right-continuous inverse, defined as in Chapter I by

$$F^{-1}(t) = \inf\{s : F(s) > t\}, \quad 0 < t < 1. \quad (3.2)$$

The representation is achieved by letting $\Omega = [0, 1]$ with Lebesgue measure (the uniform probability distribution), $\tilde{X}(\omega) = F^{-1}(\omega)$ and $\tilde{X}_n(\omega) = F_n^{-1}(\omega)$, $n \geq 1$, with an arbitrary definition for $\omega = 0$ and $\omega = 1$. The proof is based on the following four basic lemmas, the first two of which have been discussed in Sections 1.3 and 1.4 of the book.

Lemma 1.5. *If F is a cdf on $\mathbb{R}$ and U is a random variable uniformly distributed on $[0, 1]$, then $F^{-1}(U)$ is a random variable with cdf F .*

Lemma 1.6. (weak convergence criterion in terms of cdf's) *Let X and X_n be real-valued random variables with cdf's F and F_n for $n \geq 1$. Then $X_n \Rightarrow X$ as $n \rightarrow \infty$ if and only if $F_n(t) \rightarrow F(t)$ as $n \rightarrow \infty$ for all t that are continuity points of F .*

Lemma 1.7. *Let F and F_n , $n \geq 1$, be cdf's on $\mathbb{R}$. Then $F_n(t) \rightarrow F(t)$ as $n \rightarrow \infty$ for all $t \in \mathbb{R}$ that are continuity points of F if and only if $F_n^{-1}(t) \rightarrow F^{-1}(t)$ for all $t \in (0, 1)$ that are continuity points of F^{-1} .*

Lemma 1.8. *For any cdf F on $\mathbb{R}$, the set of discontinuities of F^{-1} in (3.2) is at most countably infinite.*

1.3.2. Proof for Complete Separable Metric Sspaces

The proof of Theorem 1.3.1 will be based on constructing a special family of subsets of (S, m) and relating these subsets to associated subintervals of the interval $[0, 1]$. The length of the subinterval in $[0, 1]$ (probability with respect to Lebesgue measure) will match the probability of the corresponding subset of S . The proof is a combination of Lemma 1.9 below, which shows the existence of the subsets with the required properties, and Lemma 1.10 below, which shows how to exploit such subsets to establish the Skorohod representation. Lemma 1.9 uses the separability; Lemma 1.10 uses the completeness.

A *partition* of a set A is a collection of disjoint subsets of A whose union is A . A *nested family of countable partitions* of a set A is a collection of subsets $A_{i_1, \dots, i_k}$ of A indexed by k -tuples of positive integers such that $\{A_i : i \geq 1\}$ is a partition of A and $\{A_{i_1, \dots, i_{k+1}} : i_{k+1} \geq 1\}$ is a partition of $A_{i_1, \dots, i_k}$ for all $k \geq 1$ and $(i_1, \dots, i_k) \in \mathbb{N}_+^k$. We allow $A_{i_1, \dots, i_k}$ to be empty for some $(i_1, \dots, i_k)$. For each $x \in A$, there is one and only one sequence $\{i_k : k \geq 1\}$ such that $x \in A_{i_1, \dots, i_k}$ for all k .

Example 1.1. Suppose that $S = \mathbb{R}^+$. We can obtain a nested family of countably partitions of S by letting A_i be $[i-1, i)$ and $A_{i_1, \dots, i_k}$ be the set of all positive numbers with decimal expansion beginning $(i_1-1).(i_2-1), (i_3-1), \dots, (i_k-1)$. Let $A_{i_1, \dots, i_k} = \emptyset$ if $i_j > 10$ for any $j \geq 2$. ■

We say that the *radius* of a set A in S is less than r , and write $\text{rad}(A) < r$ if $A \subseteq B_m(x, r)$ for some $x \in S$, where $B_m(x, r)$ is the open ball of radius r about x in (S, m) . As before, let ∂A be the boundary of A .

Lemma 1.9. *If P is a probability measure on a separable metric space (S, m) , then there exists a nested family of countably partitions $\{S_{i_1, \dots, i_k}\}$ of S such that, for all k and $(i_1, \dots, i_k)$,*

$$(i) \quad \text{rad}(S_{i_1, \dots, i_k}) < 2^{-k} \quad (3.3)$$

and

$$(ii) \quad P(\partial S_{i_1, \dots, i_k}) = 0. \quad (3.4)$$

Proof. Since (S, m) is a separable metric space, there exists a countable dense subset, which we can express as a sequence $\{x_i : i \geq 1\}$. For each k , we can choose an r_k such that $2^{-(k+1)} < r_k < 2^{-k}$ and

$$P(\partial B_m(x_i, r_k)) = 0 \quad \text{for all } i, \quad (3.5)$$

because there are at most countably many (r, i) such that $P(\partial B_m(x_i, r) > 0)$. Now write

$$D_i^k = B_m(x_i, r_k) - \bigcup_{j=1}^{i-1} B_m(x_j, r_k) \quad (3.6)$$

and

$$S_{i_1, \dots, i_k} = D_{i_1}^1 \cap D_{i_2}^2 \cap \dots \cap D_{i_k}^k. \quad (3.7)$$

Since

$$S_{i_1, \dots, i_k} \subseteq D_{i_k}^k \subseteq B_m(x_{i_k}, r_k) \subseteq B_m(x_{i_k}, 2^{-k}), \quad (3.8)$$

(3.3) holds. Since

$$\partial D_i^k \subseteq \bigcup_{j=1}^i \partial B_m(x_j, r_k) \quad (3.9)$$

and

$$\partial S_{i_1, \dots, i_k} \subseteq \partial D_{i_1}^1 \cup \dots \cup \partial D_{i_k}^k \subseteq \bigcup_{j=1}^k \bigcup_{l=1}^{i_j} \partial B_m(x_l, r_j), \quad (3.10)$$

(3.5) implies that (3.4) holds. ■

Lemma 1.10. *Suppose that P_0 is a probability measure on a complete metric space (S, m) with a nested family of countable partitions $\{S_{i_1, \dots, i_k}\}$ satisfying (3.3) and (3.4). If $P_n \Rightarrow P_0$ as $n \rightarrow \infty$ on (S, m) , then there exist $\tilde{X}_n$, $n \geq 0$, defined on $[0, 1]$ with Lebesgue measure, denoted by P , such that $P\tilde{X}_n^{-1} = P_n$, $n \geq 0$, and*

$$P\left(\lim_{n \rightarrow \infty} \tilde{X}_n = \tilde{X}_0\right) = 1. \quad (3.11)$$

Proof. We construct nested sequences of countable partitions of $[0, 1]$ corresponding to the given nested sequence $\{S_{i_1, \dots, i_k}\}$ of (S, m) . For $n \geq 0$, we construct subintervals $I_{i_1, \dots, i_k}^n$ corresponding to X_n . We make each subinterval closed on the left and open on the right. Let $I_1^n = [0, P_n(S_1))$ and

$$I_i^n = \left[\sum_{j=1}^{i-1} P_n(S_j), \sum_{j=1}^i P_n(S_j) \right), \quad i > 1. \quad (3.12)$$

Let $\{I_{i_1, \dots, i_{k+1}}^n : i_{k+1} \geq 1\}$ be a countable partition of subintervals of $I_{i_1, \dots, i_k}^n$. If $I_{i_1, \dots, i_k}^n = [a_n, b_n)$, then

$$I_{i_1, \dots, i_{k+1}}^n = \left[a_n + \sum_{j=1}^{i_{k+1}-1} P_n(S_{i_1, \dots, i_k, j}), a_n + \sum_{j=1}^{i_{k+1}} P_n(S_{i_1, \dots, i_k, j}) \right). \quad (3.13)$$

The length of each subinterval $I_{i_1, \dots, i_k}^n$ is the probability $P_n(S_{i_1, \dots, i_k})$. Now from each nonempty subset $S_{i_1, \dots, i_k}$ we choose one point $x_{i_1, \dots, i_k}$. For each $n \geq 0$ and $k \geq 1$, we define functions $x_n^k : [0, 1] \rightarrow S$ by letting $x_n^k(\omega) = x_{i_1, \dots, i_k}$ for $\omega \in I_{i_1, \dots, i_k}^n$. By the nested partition property and (3.3),

$$m(x_n^k(\omega), x_n^{k+j}(\omega)) < 2^{-k} \quad \text{for all } j, k, n \quad (3.14)$$

and $\omega \in [0, 1]$. Since (S, m) is a complete metric space, (3.14) implies that there is $x_n \in S$ for all $n \geq 0$ such that

$$m(x_n^k(\omega), x_n(\omega)) \rightarrow 0 \quad \text{as } k \rightarrow \infty. \quad (3.15)$$

We let $\tilde{X}_n = x_n$ on $[0, 1]$ for $n \geq 0$. Since $P_n \Rightarrow P_0$ as $n \rightarrow \infty$, $P_n(A) \rightarrow P_0(A)$ as $n \rightarrow \infty$ for all A for which $P_0(\partial A) = 0$ by Theorem 11.3.1 of the book. Hence, $P_n(S_{i_1, \dots, i_k}) \rightarrow P_0(S_{i_1, \dots, i_k})$ by (3.4). Consequently, the length of the intervals $I_{i_1, \dots, i_k}^n$ converge to the length of the intervals $I_{i_1, \dots, i_k}^0$.

as $n \rightarrow \infty$. Since

$$\begin{aligned} m(\tilde{X}_n(\omega), \tilde{X}_0(\omega)) &\leq m(\tilde{X}_n(\omega), x_n^k(\omega)) + m(x_n^k(\omega), x_0^k(\omega)) \\ &\quad + m(x_0^k(\omega), \tilde{X}_0(\omega)) \\ &\leq 2^{-(k-1)} + m(x_n^k(\omega), x_0^k(\omega)) , \end{aligned} \quad (3.16)$$

for all ω in the interior of $I_{i_1, \dots, i_k}^0$,

$$\lim_{n \rightarrow \infty} m(\tilde{X}_n(\omega), \tilde{X}_0(\omega)) \leq 2^{-(k-1)} . \quad (3.17)$$

Since k is arbitrary, we must have $\tilde{X}_n(\omega) \rightarrow \tilde{X}_0(\omega)$ as $n \rightarrow \infty$ for all but at most countably many $\omega \in [0, \infty)$.

It remains to show that $\tilde{X}_n$ has the probability law P_n for $n \geq 0$. It suffices to show that $P(\tilde{X}_n \in A) = P_n(A)$ for each A such that $P_n(\partial A) = 0$. Let A be such a set. Let A^k be the union of the sets $S_{i_1, \dots, i_k}$ such that $S_{i_1, \dots, i_k} \subseteq A$ and let A'^k be the union of the sets $S_{i_1, \dots, i_k}$ such that $S_{i_1, \dots, i_k} \cap A \neq \emptyset$. Then $A^k \subseteq A \subseteq A'^k$ and, by construction above,

$$P(\tilde{X}_n \in A^k) = P_n(A^k) \quad \text{and} \quad P(\tilde{X}_n \in A'^k) = P_n(A'^k) . \quad (3.18)$$

Now let

$$C^k = \{x \in S : m(x, \partial A) \leq 2^{-k}\} . \quad (3.19)$$

Then $A'^k - A^k \subseteq C^k \downarrow \partial A$ as $k \rightarrow \infty$. Since $P_n(\partial A) = 0$ by assumption, $P_n(C^k) \downarrow 0$ as $k \rightarrow \infty$. Hence

$$P(\tilde{X}_n \in A) = \lim_{k \rightarrow \infty} P(\tilde{X}_n \in A^k) = \lim_{k \rightarrow \infty} P_n(A^k) = P_n(A) . \quad \blacksquare \quad (3.20)$$

1.3.3. Proof for Separable Metric Spaces

We now do the proof of Theorem 1.3.1 without assuming completeness. Start by letting P_n be the probability distribution of X_n on S for $n \geq 0$. Let the underlying probability space be the product space $\Omega \equiv S^\infty$ with elements $\omega \equiv \{s_k : k \geq 0\}$. Let $\tilde{X}_n$ be the coordinate mapping, e.g., $\tilde{X}_n(\{s_k : k \geq 0\}) = s_n$, $n \geq 0$. To quickly get the idea, first suppose that $P_n(\{s\}) = 1$ for all $n \geq 0$. In this special case we can let the probability measure P on Ω be the product measure $P = \delta_s \times \delta_s \times \dots$, where δ_s is the Dirac measure assigning probability 1 to the point $s \in S$. Then P assigns probability 1 to the sequence $\{s_n : n \geq 0\}$ where $s_n = s$ for all n . Since $P(\tilde{X}_n = s) = 1$ for all n ,

$$P(\tilde{X}_n = \tilde{X}_0 \quad \text{for all } n) = P\left(\bigcap_{n=0}^{\infty} \{\tilde{X}_n = s\}\right) = 1 . \quad (3.21)$$

To continue to develop the idea of the approach, now suppose that each probability measure P_n , $n \geq 0$, concentrates all probability on a common finite subset of S . Thus it suffices to assume that S is finite. For a sequence $\{k_n : n \geq 1\}$ with $k_n \rightarrow \infty$ as $n \rightarrow \infty$ to be defined later, let

$$U_k = \cap_{n:k_n \geq k} \{\tilde{X}_n = \tilde{X}_0\} . \quad (3.22)$$

(Note that we have a strong form of convergence on U_k .) Also let $\{Q_n : n \geq 1\}$ be a sequence of probability measures on S to be defined later. Now let $P_{j,s}$ be the product measure

$$P_{j,s} = \delta_s \times \prod_{n=1}^{\infty} P_{j,s,n} , \quad (3.23)$$

where $P_{j,s,n}$ is a probability measure on S defined by

$$P_{j,s,n} = \begin{cases} Q_n & \text{if } 0 \leq k_n < j \\ \delta_s & \text{if } j \leq k_n \leq \infty . \end{cases} \quad (3.24)$$

Then let P'_j be a mixture of the probabilities $P_{j,s}$ in (3.23) with respect to P_0 , in particular,

$$P'_j = \sum_{s \in S} P_0(\{s\}) P_{j,s} . \quad (3.25)$$

Next let $\{w_k : k \geq 1\}$ and $\{q_k : k \geq 0\}$ be sequences of numbers with

$$w_k \geq 0, \quad \sum_{k=1}^{\infty} w_k = 1, \quad q_0 = 0, \quad q_k = \sum_{j=1}^k w_j < 1, \quad 1 \leq k < \infty . \quad (3.26)$$

Then let P be a mixture of the probabilities P'_j in (3.25) using the weights w_j in (3.26), i.e.,

$$P = \sum_{j=1}^{\infty} w_j P'_j . \quad (3.27)$$

We will show that this construction does the job with an appropriate choice of the sequences $\{k_n : n \geq 1\}$ and $\{Q_n : n \geq 1\}$. (The weights w_k in (3.26) can be arbitrary subject to the conditions in (3.26).)

Note that $P_{j,s}$ in (3.23) attaches positive probability only to sets of sequences $\{s_n : n \geq 0\}$ such that $s_n = s$ for all but a finite number of n (those n for which $0 \leq k_n < j$). Thus even though S^∞ is uncountably infinite, $P_{j,s}$ has finite support. Since S is finite, P'_j in (3.25) also has finite support. All

sequences $\{s_n : n \geq 0\}$ in S^∞ with positive P -measure have $s_n = s$ for all sufficiently large n for some s .

By (3.23), $P_{j,s}(\tilde{X}_0 = s) = 1$. Thus, by (3.25) and (3.27), $P(\tilde{X}_0 = s) = P'_j(\tilde{X}_0 = s) = P_0(\{s\})$ for all $s \in S$. Hence $P\tilde{X}_0^{-1} = P_0$ or, equivalently, $\tilde{X}_0 \stackrel{d}{=} X_0$.

Next $P_{j,s}(\tilde{X}_n = s) = P_{j,s,n}(\{s\})$ for $n \geq 1$. Note that $P_{j,s}(U_k) = 1$ for $j \leq k$, where U_k is given in (3.22), so that $P'_j(U_k) = 1$ if $j \leq k$ and $P(U_k) \geq q_k$. Since $q_k \rightarrow 1$ as $k \rightarrow \infty$ by (3.26), $\tilde{X}_n \rightarrow \tilde{X}_0$ as $n \rightarrow \infty$ almost uniformly on Ω with respect to P , i.e., for any $\epsilon > 0$, there exists a subset U_k of S with $P(U_k) > 1 - \epsilon$ such that $\tilde{X}_n$ converges uniformly to $\tilde{X}_0$ as $n \rightarrow \infty$ on U_k . (In our finite-state-space setting, we actually have $\tilde{X}_n = \tilde{X}_0$ on U_k for all n such that $k_n \geq k_0$ by (3.22) and (3.26).) For ϵ given, choose k so that $q_k > 1 - \epsilon$. By Egoroff's theorem, p. 89 of Halmos (1956), that implies that

$$P\left(\lim_{n \rightarrow \infty} \tilde{X}_n = \tilde{X}_0\right) = 1. \quad (3.28)$$

The difficult part is to obtain $\tilde{X}_n \stackrel{d}{=} X_n$ for $n \geq 1$. The construction above yields

$$P(\tilde{X}_n = s) = q_{k_n} P_0(\{s\}) + (1 - q_{k_n}) Q_n(\{s\}) \quad \text{for all } n. \quad (3.29)$$

We now choose the sequences $\{k_n : n \geq 1\}$ and $\{Q_k : k \geq 1\}$ to achieve $P\tilde{X}_n^{-1} = P_n$ for all n . Note that (3.29) is equivalent to

$$Q_n(\{s\}) = P_n(\{s\}) + \frac{q_{k_n}}{1 - q_{k_n}} (P_n(\{s\}) - P_0(\{s\})) \quad (3.30)$$

provided $k_n < \infty$. If $k_n = \infty$, then $q_{k_n} = 1$, so that we must have $P_n(\{s\}) = P_0(\{s\})$, and then any $Q_n(\{s\})$ will do.

Thus, let

$$Q(k, s, n) \equiv P_n(\{s\}) + \frac{q_k}{1 - q_k} (P_n(\{s\}) - P_0(\{s\})), \quad (3.31)$$

$$m_{k,n} \equiv \min_{s \in S} Q(k, s, n), \quad (3.32)$$

$$k_n = \sup\{j \geq 0 : m_{j,n} \geq 0\} \quad (3.33)$$

and

$$Q_n(\{s\}) = Q(k_n, s, n) \quad \text{for } k_n < \infty. \quad (3.34)$$

Note that $m_{0,n} \geq 0$, so that $k_n \leq \infty$ is well defined in (3.33). Note that $\sum_{s \in S} Q(k, s, n) = 1$ for all k , $0 \leq k < \infty$, and $Q(k_n, s, n) \geq 0$ by (3.32)

and (3.33). Thus, under (3.31)–(3.34), Q_n is a probability measure on S satisfying (3.29) provided that $k_n < \infty$.

Since $\sum_{s \in S} Q(k, s, n) = 1$, we must have $0 \leq Q(k, s, n) \leq 1$ for $Q(k, s, n)$ in (3.31). Since $q_k \rightarrow 1$ as $k \rightarrow \infty$, $q_k/(1 - q_k) \rightarrow \infty$ as $k \rightarrow \infty$. Hence, we must have $P_n(\{s\}) = P_0(\{s\})$ for all n if $k_n = \infty$, under which (3.29) has been shown to hold for any probability measure Q_n .

We now show that $k_n \rightarrow \infty$ as $n \rightarrow \infty$. Since $P_n(\{s\}) \rightarrow P_0(\{s\})$ as $n \rightarrow \infty$ for each s , $Q(k, s, n) \rightarrow P_0(\{s\})$ as $n \rightarrow \infty$ for each s and k , $1 \leq k < \infty$. This, together with the fact that $Q(k, s, n) \geq 0$ if $P_0(\{s\}) = 0$, implies that $Q(k, s, n)$ is ultimately nonnegative for all sufficiently large n depending upon k . Thus, for each k , there is an index n_k such that $Q(k, s, n) \geq 0$, and thus $m_{k,n} \geq 0$, for all $n \geq n_k$. Since $m_{k,n} \geq 0$ implies $k_n \geq k$, we can conclude that, for all $n \geq n_k$, $k_n \geq k$. Hence, $k_n \rightarrow \infty$ as $n \rightarrow \infty$.

We now turn to the general case: We now assume that S is a separable metric space. We start by constructing a finite collection of subsets appropriately approximating S . This step is a minor modification of Lemma 1.9.

Lemma 1.11. *If P is a probability measure on separable metric space (S, m) , then for any $\delta, \epsilon > 0$ there exist disjoint subsets $S_{i_1, \dots, i_k}$ of S , $1 \leq i_j \leq i'_j$, $1 \leq j \leq k$, such that, for all k and $(i_1, \dots, i_k)$, (3.3) holds for $2^{-k} < \delta$, (3.4) holds and*

$$P\left(\bigcup_{i_1=1}^{i'_1} \cdots \bigcup_{i_k=1}^{i'_k} S_{i_1, \dots, i_k}\right) > 1 - \epsilon. \quad (3.35)$$

Proof. We use the construction in Lemma 1.9. Choose i'_1 such that $P(S_1 \cup \cdots \cup S_{i'_1}) > 1 - \epsilon 2^{-1}$; choose i'_2 such that

$$P(S_{i_1, 1} \cup \cdots \cup S_{i_1, i'_2}) > 1 - P(S_{i_1})\epsilon 2^{-2} \quad (3.36)$$

for all i_1 , $1 \leq i_1 \leq i'_1$; choose i'_{j+1} such that

$$P(S_{i_1, \dots, i_j, 1} \cup \cdots \cup S_{i_1, \dots, i_j, i'_{j+1}}) > 1 - P(S_{i_1, \dots, i_j})\epsilon 2^{-j} \quad (3.37)$$

for all $(i_1, \dots, i_j) \leq (i'_1, \dots, i'_j)$. Stop at k with $2^{-k} < \delta$, so that (3.3) holds. Then

$$P\left(\bigcup_{i_1=1}^{i'_1} \cdots \bigcup_{i_k=1}^{i'_k} S_{i_1, \dots, i_k}\right) > 1 - \epsilon(2^{-1} + \cdots + 2^{-k}) > 1 - \epsilon, \quad (3.38)$$

so that (3.35) holds. ■

We now return to the proof of the theorem. Let $\{\delta_k : k \geq 1\}$ and $\{\epsilon_k : k \geq 1\}$ be sequences of positive numbers such that $\delta_k \rightarrow 0$, $\epsilon_k \rightarrow 0$ and $\sum_{k=1}^{\infty} \epsilon_k < \infty$. For each k , let $\{C_{k,j} : 0 \leq j \leq n_k\}$ be the finite collection of subsets $S_{i_1, \dots, i_k}$ in Lemma 1.11 constructed with respect to P_0 , where δ and ϵ for k are required to be δ_k and ϵ_k . Let $C_{k,0} = S - \cup_{j=1}^{n_k} C_{k,j}$. By (3.35), $P_0(C_{k,0}) < \epsilon_k$.

With $\tilde{X}_n$ the coordinate projections on S^∞ as before, instead of (3.22), let

$$U_k = \cap_{n: k_n \geq k} \{m(\tilde{X}_n, \tilde{X}_0) \leq \delta_{k_n}\} \quad (3.39)$$

where $\delta_\infty = 0$. (The separability of (S, m) is used to have $\{m(\tilde{X}_n, \tilde{X}_0) \leq \delta_{k_n}\}$ and thus U_k be measurable.) Given that $k_n \rightarrow \infty$ as $n \rightarrow \infty$, $\tilde{X}_n \rightarrow \tilde{X}_0$ uniformly on U_k . To apply Egoroff's theorem, we will need to show that $P(U_k) \rightarrow 1$ as $k \rightarrow \infty$.

Let Π_k be the collection of sets $C_{k,j}$, $1 \leq j \leq n_k$, and let $\Pi_0 = S$. We now modify the finite-state-space proof above, letting $C_{k,j}$ play the role of s . Let the weights w_k and their partial sums q_k be defined by (3.26). Paralleling (3.31)–(3.34), for $0 \leq k < \infty$, let

$$Q(k, C, n) = P_n(C) + \frac{q_k}{1 - q_k} (P_n(C) - P_0(C)) , \quad (3.40)$$

$$m_{k,n} = \min_{C \in \Pi_k} \{Q(k, C, n)\} , \quad (3.41)$$

$$k_n = \sup\{j \geq 0 : m_{j,n} \geq 0\} \quad (3.42)$$

and

$$Q_n(C) = Q(k_n, C, n) . \quad (3.43)$$

Since $P_n(C) \rightarrow P_0(C)$ as $n \rightarrow \infty$ for all $C \in \Pi_k$, $k_n \rightarrow \infty$ as $n \rightarrow \infty$ by the same argument as before.

Paralleling (3.23), let $P_{j,s}$ be the product measure

$$P_{j,s} = \delta_s \times \prod_{n=1}^{\infty} P_{j,s,n} , \quad (3.44)$$

where $P_{j,s,n}$ is a probability measure on S defined by

$$P_{j,s,n} = \begin{cases} Q_n & \text{if } 0 \leq k_n < j \\ P_n(\cdot | C_{k_n,s}) & \text{if } j \leq k_n < \infty \\ \delta_s & \text{if } k_n = \infty , \end{cases} \quad (3.45)$$

where $P_n(\cdot | C_{k_n,s})$ is the conditional probability measure with $C_{k_n,s}$ being the element of Π_k containing $s \in S$. Note that $P_{j,s,n}$ in (3.45) has three

possibilities instead of only the two in (3.24). Unlike the case of finite S , $P_{j,s}$ in (3.44) does not have finite support, but if $s \in C_{k_n,i}$, then $P_{j,s}$ has support on the set of sequences $\{s_n : n \geq 0\}$ such that $s_n \in C_{k_n,i}$ for all but finitely many n , in particular, for all n such that $k_n \geq j$. On this subset of sequences, $m(\tilde{X}_n, \tilde{X}_0) \leq \delta_{k_n}$ for all n such that $k_n \geq j$.

Paralleling (3.25), let

$$P'_j(A) = \int_S P_0(ds) P_{j,s}(A) . \quad (3.46)$$

The integral in (3.46) is well defined since $P_{j,s}(A)$ is a measurable function on S for each A a cylinder set with finite base in the σ -field on S^∞ ; see pp. 74–76 of Neveu (1965). Note that P'_j has support on the set of sequences $\{s_n : n \geq 0\}$ such that $s_n \in C_{k_n,i}$ for all but finitely many n , for some i . Thus

$$P'_j(m(\tilde{X}_n, \tilde{X}_0) \leq \delta_{k_n}) \geq 1 - P(C_{k_n,0}) > 1 - \epsilon_{k_n} . \quad (3.47)$$

Paralleling (3.27), let

$$P = \sum_{j=1}^{\infty} w_j P'_j . \quad (3.48)$$

As before, the construction yields $P\tilde{X}_0^{-1} = P_0$. The probability distribution of $\tilde{X}_n$ is

$$P\tilde{X}_n^{-1} = \begin{cases} q_{k_n} \sum_{C \in \Pi_{k_n}} P_n(\cdot|C) P_0(C) + (1 - q_{k_n}) Q_n & \text{if } k_n < \infty \\ P_0 & \text{if } k_n = \infty . \end{cases} \quad (3.49)$$

For n such that $k_n < \infty$, let

$$Q_n = \sum_{C \in \Pi_k} Q(k_n, C, n) P_n(\cdot|C) . \quad (3.50)$$

Combining (3.40), (3.49) and (3.50), we see that $P\tilde{X}_n^{-1} = P_n$ if $k_n < \infty$. On the other hand, as before, if $k_n = \infty$, then we are forced to have $P_n(C) = P_0(C)$ for all $C \in \Pi_k$ for any $k \geq 1$, but that implies that $P_n = P_0$. (We can apply the reasoning in the proof of Lemma 1.10 using (3.18) and (3.19).)

Finally, it remains to show that $P(U_k) \rightarrow 1$ as $k \rightarrow \infty$ for U_k in (3.39). However,

$$\begin{aligned} P(U_k) &= \sum_{j=1}^{\infty} w_j P'_j(U_k) \geq \sum_{j=1}^k w_j P'_j(U_k) \\ &\geq \left(\sum_{j=1}^k w_k \right) \left(1 - \sum_{j=k}^{\infty} \epsilon_j \right) \rightarrow 1 \quad \text{as } k \rightarrow \infty , \end{aligned} \quad (3.51)$$

since, for $j \leq k \leq k_n$,

$$1 - P'_j(U_k) \leq P_0(\cup_{l=k}^{\infty} C_{l,0}) \leq \sum_{l=k}^{\infty} \epsilon_l, \quad (3.52)$$

because $P'_{j,s}$ assigns probability 1 to product sets in which all coordinates are in common sets $C_{i,k}$.

1.4. The “Weak” in Weak Convergence

This section is devoted, not to a proof of a theorem, but to an explanation of a term – the adjective “weak” in “weak convergence.” The term “weak” can be understood from a Banach-space perspective.

The starting point is the definition of convergence $P_n \Rightarrow P$; i.e., $P_n \Rightarrow P$, if

$$\lim_{n \rightarrow \infty} \int_S f dP_n = \int_S f dP \quad (4.1)$$

for all functions f in $C(S)$, the space of all continuous bounded real-valued functions on S .

The space $C(S)$ of continuous bounded real-valued functions h on S used in definition (4.1) is a Banach space (a complete normed linear topological space) with the uniform norm

$$\|h\| \equiv \sup_{s \in S} |h(s)|.$$

The adjoint or conjugate space of $C(S)$, the space of all continuous linear real-valued functions L on $C(S)$, denoted by $C^*(S)$, turns out to be the space $Z(S)$ of all finite signed measures μ on S , defined via

$$L(h) \equiv \int_S h d\mu;$$

e.g., see pp. 262, 419 of Dunford and Schwartz (1958) or Chapter 9 of Simmons (1963).

The adjoint space B^* of any Banach space B is itself a Banach space with the norm

$$\|L\| \equiv \sup\{\|L(b)\| : b \in B, \|b\| \leq 1\}.$$

Since B^* is a Banach space, one can consider its adjoint space B^{**} . There is a natural embedding of B in B^{**} so that we can regard B as a subset of

B^{**} . (Just let $L_b(f) = f(b)$ for $b \in B$ and $f \in B^*$.) When $B = B^{**}$, B is said to be *reflexive*. However, $C(S)$ is reflexive only when S is finite. So, in our setting with infinite S , $C(S)$ is a proper subset of $C^{**}(S)$.

Instead of the topology induced on a Banach space B by its norm, it is sometimes of interest to consider a weaker topology on B called the *weak topology*, which is the weakest topology such that all the functions in B^* remain continuous; i.e., $b_n \rightarrow b$ in B with the weak topology if and only if $L(b_n) \rightarrow L(b)$ for all L in B^* . Furthermore, on the adjoint space B^* one can also consider a still weaker topology called the *weak* topology*, which is the weakest topology such that all the functions in B , regarded as a subset of B^{**} , remain continuous. Thus the weak* topology on $Z(S) = C^*(S)$ relativized to the subset $P(S)$ is what is characterized by (4.1). (The discussion also implies that the weak topology on $Z(S)$ is stronger than the weak* topology on $Z(S)$, so the terminology “weak convergence” is something of a misnomer. From this Banach-space perspective, we should actually call weak convergence $P_n \Rightarrow P$ *weak* convergence*.) ■

1.5. Continuous-Mapping Theorems

In this section we supplement the discussion of the continuous-mapping approach in Section 3.4 of the book by providing proofs for the unproved theorems. We first prove the Lipschitz mapping theorem, which comes from Whitt (1974).

1.5.1. Proof of the Lipschitz Mapping Theorem

We now prove the Lipschitz mapping theorem, Theorem 2.4.2 in the book. First suppose that (S, m) is a separable metric space and $B = S$. Then we can employ the Strassen representation theorem, Theorem 11.3.5 in the book. It is elementary that the Lipschitz property is inherited by the in-probability distance p : Given $P(m(X, Y) > \delta) < \delta$, the Lipschitz property of g implies that $P(m'(g(X), g(Y)) > K\delta) < \delta$, so that $p(g(X), g(Y)) \leq (K \vee 1)p(X, Y)$. By the Strassen representation theorem, for X, Y and positive ϵ given, we can find $\tilde{X}, \tilde{Y}$ on a common probability space so that $\tilde{X} \stackrel{d}{=} X$, $\tilde{Y} \stackrel{d}{=} Y$ and

$$p(\tilde{X}, \tilde{Y}) \leq \pi(X, Y) + \epsilon.$$

Hence,

$$\pi(g(X), g(Y)) = \pi(g(\tilde{X}), g(\tilde{Y})) \leq p(g(\tilde{X}), g(\tilde{Y}))$$

and

$$p(g(\tilde{X}), g(\tilde{Y})) \leq (K \vee 1)p(\tilde{X}, \tilde{Y}) \leq (K \vee 1)(\pi(X, Y) + \epsilon) .$$

Since ϵ was arbitrary, we have the desired conclusion.

Now we consider the general case, for which we argue directly. Let B be the subset for which $P(Y \in B) = 1$. The Lipschitz property implies that

$$B \cap g^{-1}(A)^\delta \subseteq g^{-1}(A^\epsilon) \quad \text{in } S \quad \text{for } \delta \leq \epsilon/K$$

and any $A \in \mathcal{S}'$. Hence,

$$\begin{aligned} & \pi(g(X), g(Y)) \\ &= \inf \{ \epsilon > 0 : P(g(X) \in A) \leq \epsilon + P(g(Y) \in A^\epsilon) \text{ for all } A \in \mathcal{S}' \} \\ &= \inf \{ \epsilon > 0 : P(X \in g^{-1}(A)) \leq \epsilon + P(Y \in g^{-1}(A^\epsilon)) \text{ for all } A \in \mathcal{S}' \} \\ &\leq \inf \{ \epsilon > 0 : P(X \in g^{-1}(A)) \leq \epsilon + P(Y \in B \cap g^{-1}(A)^\delta) \text{ for all } A \in \mathcal{S}' \} \\ &\leq \inf \{ \epsilon > 0 : P(X \in g^{-1}(A)) \leq \epsilon + P(Y \in g^{-1}(A)^\delta) \text{ for all } A \in \mathcal{S}' \} \\ &\leq \inf \{ \epsilon > 0 : P(X \in A) \leq \epsilon + P(Y \in A^\delta) \text{ for all } A \in \mathcal{S} \} \\ &\leq (1 \vee K)\pi(X, Y) . \quad \blacksquare \end{aligned}$$

Example 1.5.1. *The advantage of the Prohorov metric on $\mathcal{P}(\mathbb{R})$.* Even on the real line $\mathbb{R}$, the Prohorov metric is useful to establish rate of convergence results, because the Lipschitz mapping theorem does not apply to two other metrics commonly used. On $\mathcal{P}(\mathbb{R})$ one often uses the *Lévy metric* λ , which is defined just as the Prohorov metric π in (2.2) except that only sets of the form $A = (-\infty, x]$ are used. The *uniform metric for cdf's* is also sometimes used; i.e.,

$$\|F_1 - F_2\| \equiv \mu(P_1, P_2) \equiv \sup\{|P_1(A) - P_2(A)| : A = (-\infty, x]\} ,$$

where $F_i(x) \equiv P((-\infty, x])$. The uniform-cdf metric μ also induces weak convergence at limiting probability measures without atoms. However, the Lipschitz theorem is not valid for λ and μ . To see that, for $n \geq 1$, let

$$P(X_n = 2j) = P(Y_n = 2j + 1) = 1/n \quad \text{for } 1 \leq j \leq n ,$$

and let $g : \mathbb{R} \rightarrow \mathbb{R}$ be defined by

$$g(t) = \sin(\pi t/2) \quad \text{for } t \in \mathbb{R} .$$

Clearly, g is Lipschitz with Lipschitz constant 1, but

$$\lambda(X_n, Y_n) \leq \mu(X_n, Y_n) = 1/n ,$$

while

$$P(g(X_n) = 0) = P(g(Y_n) \in \{-1, 1\}) = 1 \quad \text{for all } n ,$$

so that

$$\mu(g(X_n), g(Y_n)) \geq \lambda(g(X_n), g(Y_n)) = 1/2 \quad \text{for all } n .$$

Given a bound with the Prohorov metric π in $\mathcal{P}(\mathbb{R})$, we can obtain corresponding bounds with the metrics λ and μ . First we use the inequality $\lambda \leq \pi$. In many cases we can relate λ and μ : When a probability measure P_1 on $\mathbb{R}$ has a Lipschitz cdf F_1 with Lipschitz constant c , i.e., when

$$|F_1(t_1) - F_1(t_2)| \leq c|t_1 - t_2| ,$$

then we have the ordering

$$\mu(P_1, P_2) \leq (1 + c)\lambda(P_1, P_2) \quad \text{for all } P_2 \in \mathcal{P}(\mathbb{R}) . \quad \blacksquare \quad (5.1)$$

1.5.2. Proof of the Continuous-Mapping Theorems

We now turn to Theorem 3.4.3 of the book, following Billingsley (1968, Section 5), which we restate here. Let $Disc(g)$ be the set of discontinuity points of the function g .

Theorem 1.5.1. (continuous-mapping theorem) *If $X_n \Rightarrow X$ in (S, m) and $g : (S, m) \rightarrow (S', m')$ is measurable with $P(X \in Disc(g)) = 0$, then $g(X_n) \Rightarrow g(X)$.*

We first establish the measurability of $Disc(g)$ (even if g is not measurable).

Lemma 1.5.1. (measurability of the set of discontinuity points) *For $g : (S, m) \rightarrow (S', m')$, $Disc(g) \in \mathcal{S}$.*

Proof. For any $y, z \in S$ with $m'(g(y), g(z)) \geq \epsilon$ and $\epsilon > 0$, let

$$A_{\epsilon, \delta}(y, z) \equiv \{x \in S : m(x, y) < \delta \text{ and } m(x, z) < \delta\} .$$

Then the complement is

$$A_{\epsilon, \delta}^c(y, z) \equiv \{x \in S : m(x, y) \geq \delta \text{ or } m(x, z) \geq \delta\} .$$

It is easy to see that $A_{\epsilon,\delta}^c(y, z)$ is closed, so that $A_{\epsilon,\delta}(y, z)$ is open, as necessarily is

$$A_{\epsilon,\delta} \equiv \bigcup_y \bigcup_z A_{\epsilon,\delta}(y, z) .$$

Since

$$Disc(g) = \bigcup_{\epsilon} \bigcap_{\delta} A_{\epsilon,\delta} ,$$

where ϵ and δ run over the positive rationals, $Disc(g)$ is a $G_{\delta\sigma}$, implying that $Disc(g) \in \mathcal{S}$. ■

Proof of Theorem 1.5.1. By Theorem 11.3.4 (iii) in the book, it suffices to show that

$$\overline{\lim}_{n \rightarrow \infty} P(g(X_n) \in F) \leq P(g(X) \in F)$$

for each closed subset $F \in \mathcal{S}'$. Given that $X_n \Rightarrow X$, we have

$$\begin{aligned} \overline{\lim}_{n \rightarrow \infty} P(g(X_n) \in F) &= \overline{\lim}_{n \rightarrow \infty} P(X_n \in g^{-1}(F)) \\ &\leq \overline{\lim}_{n \rightarrow \infty} P(X_n \in g^{-1}(F)^-) \\ &\leq P(X \in g^{-1}(F)^-) . \end{aligned}$$

However, $P(X \in g^{-1}(F)^-) = P(X \in g^{-1}(F))$ because $P(Disc(g)) = 0$ and $g^{-1}(F)^- \subseteq g^{-1}(F) \cup Disc(g)$. ■

Finally, we treat Theorem 3.4.4 of the book, involving a sequence of measurable mappings:

Theorem 1.5.2. (generalized continuous-mapping theorem) *Let g and g_n , $n \geq 1$, be measurable functions mapping (S, m) into (S', m') . Let the range (S', m') be separable. Let E be the set of x in S such that $g_n(x_n) \rightarrow g(x)$ fails for some sequence $\{x_n : n \geq 1\}$ with $x_n \rightarrow x$ in S . If $X_n \Rightarrow X$ in (S, m) and $P(X \in E) = 0$, then $g_n(X_n) \Rightarrow g(X)$ in (S', m') .*

Here we need to assume that the range is a separable metric space. Again we follow Billingsley (1968, Section 5).

Lemma 1.5.2. (measurability of the bad set) *Suppose that g_n , $n \geq 1$, and g are measurable functions from a metric space (S, m) into a separable metric space (S', m') . Let E be the set of x in S such that $g_n(x_n) \rightarrow g(x)$ fails for some sequence $\{x_n : n \geq 1\}$ with $m(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$. Then E is a measurable subset of S .*

Proof. Let $B_{\epsilon,\delta,i}$ be the set of x in S such that $m'(g(x), g_i(y)) \geq \epsilon$ for some y with $m(x, y) < \delta$. Note that

$$E = \cup_{\epsilon} \cap_{\delta} \cap_{k \geq 1} \cup_{i \geq k} B_{\epsilon,\delta,i} \quad (5.2)$$

where ϵ and δ range over the positive rationals. We would be done if we could conclude that $B_{\epsilon,\delta,i}$ is measurable, but we do not know that. Note that $B_{\epsilon,\delta,i}$ is decreasing in ϵ . Hence (5.2) remains true if $B_{\epsilon,\delta,i}$ is replaced by $B_{\epsilon/2,\delta,i}$. It thus suffices to show that, for all (ϵ, δ, i) , there are sets $C_{\epsilon,\delta,i} \in \mathcal{S}$ such that

$$B_{\epsilon,\delta,i} \subseteq C_{\epsilon,\delta,i} \subseteq B_{\epsilon/2,\delta,i} . \quad (5.3)$$

Since (S', m') is separable, we can find a sequence $\{u_k : k \geq 1\}$ dense in S' . Let $A_{\epsilon,k} = \{x : m'(g(x), u_k) < \epsilon/4\}$ and note that $A_{\epsilon,k} \in \mathcal{S}$ and $S = \cup_k A_{\epsilon,k}$. Then (5.3) holds if

$$C_{\epsilon,\delta,i} = \cup_k (A_{\epsilon,k} \cap J_{\epsilon,\delta,i,k}) ,$$

where $J_{\epsilon,\delta,i,k}$ is the set of x such that $m'(g_i(y), g(z)) \geq \epsilon$ for some pair of points y, z in S with $m(x, y) < \delta$, $m(x, z) < \delta$ and $z \in A_{\epsilon,k}$. It is not difficult to see that $J_{\epsilon,\delta,i,k}^c$ is closed, so that $J_{\epsilon,\delta,i,k}$ is open and $C_{\epsilon,\delta,i} \in \mathcal{S}$. ■

Proof of Theorem 1.5.2. By Lemma 1.5.2, $E \in \mathcal{S}$. From Theorem 11.3.4 (iv) in the book, it suffices to show that

$$P(g(X) \in G) \leq \varliminf_{n \rightarrow \infty} P(g_n(X_n) \in G)$$

for every open G in S' . If $x \in E^c$ and $g(x) \in G$, then there must exist k and δ such that $g_i(y) \in G$ if $i \geq k$ and $m(x, y) < \delta$, so that $x \in T_k^o$, the interior of T_k , where

$$T_k = \cap_{i \geq k} g_i^{-1}(G) .$$

Consequently,

$$g^{-1}(G) \subseteq E \bigcup \cup_{k=1}^{\infty} T_k^o .$$

Since $P(X \in E) = 0$ and $T_k^o \subseteq T_{k+1}^o$, for any given ϵ there is a k such that

$$P(X \in g^{-1}(G)) \leq P(X \in \cup_k T_k^o) \leq P(X \in T_k^o) + \epsilon$$

for $k \geq k_0$. Since $X_n \Rightarrow X$ and $T_k \subseteq g_n^{-1}(G)$ for $n \geq k$,

$$P(X \in T_k^o) \leq \varliminf_{n \rightarrow \infty} P(X_n \in T_k^o) \leq \varliminf_{n \rightarrow \infty} P(X_n \in g_n^{-1}(G)) .$$

Since ϵ was arbitrary, the proof is completed by combining these two strings of inequalities. ■

The continuous-mapping approach to stochastic-process limits leads us to focus on the underlying sample paths of the stochastic processes. Thus the continuous-mapping approach is a *sample-path method*. In recent years, probabilists have tended to favor sample-path methods over more traditional analytic methods, because they are less removed from the phenomenon under study. However, the two approaches often can be fruitfully combined.

Many traditional analytic results are based on transforms, such as the characteristic function (version of the Fourier transform), probability generating function (or z transform) and the Laplace transform, as can be seen from Feller (1971). Fortunately, the analytic approach has been applied with great success over the years to yield explicit expressions for many probability distributions of interest in the form of transforms. That is true for many of the limit processes that we will consider. Thus we can use previous analytic results to obtain explicit transforms for approximating distributions. We then can apply numerical transform inversion to compute the probability distribution itself; e.g., see Abate and Whitt (1992a, 1995), Choudhury, Lucantoni and Whitt (1994) and Abate, Choudhury and Whitt (1999).

For example, as shown in Section 8.5 of the book and Section 5.2 here, the heavy-traffic limit for a queue with heavy-tailed distributions is often a reflection of a stable Levy motion or more general Levy process without negative jumps. These limit processes are somewhat complicated, but fortunately the analytic approach has shown that the steady-state distribution has a relatively simple expression via its Laplace transform, which is known as the generalized Pollaczek-Khintchine transform. Thus we can calculate the steady-state distribution of the limit process by applying numerical transform inversion.

Chapter 2

Stochastic-Process Limits

2.1. Introduction

Chapters 4 and 7 of the book present a panorama of stochastic-process limits. In this chapter we present even more material. In Section 2.2 we present an introduction to strong approximations and the rates of convergence in the setting of Donsker's theorem that they imply using the Prohorov metric. In Section 2.3 we present additional Brownian limits under weak dependence; here we focus on Markov and regenerative structure.

In Section 2.4 we briefly discuss the convergence to general Lévy processes that holds when we have a sequence of random walks (based on a double sequence of random walk steps). Finally, in Section 2.5 we point out that the linear-process representation assumed with strong dependence in Sections 4.6 and 4.7 of the book arises naturally from modelling when we take a time-series perspective.

2.2. Strong Approximations and Rates of Convergence

In Sections 1.4 and 4.3 of the book we noted that the CLT and FCLT are invariance principles, meaning that the same limits occur in great generality. In the IID case we only need the summands X_n to have finite variance. However, the quality of the approximation for any given n is affected by the distribution of X_n . Indeed, that is obvious for the CLT: If $X_n \stackrel{d}{=} N(0, \sigma^2)$, then the limit can be replaced by equality in distribution. Moreover, the closer the distribution of X_n is to the normal distribution, the better the

normal approximation for the scaled partial sum should be. More generally, the advantage of extra structure in the distribution of X_n can be seen from more refined results giving bounds on the rate of convergence and asymptotic expansions. We review some of these results in this section.

2.2.1. Rates of Convergence in the CLT

A bound on the rate of convergence in the basic CLT, given a finite third absolute moment of a summand, is provided by the Berry-Esseen theorem; see p. 542 of Feller (1971). To state it, we use the uniform metric on cdf's, defined by

$$\|F_1 - F_2\| \equiv \sup_x |F_1(x) - F_2(x)|. \quad (2.1)$$

As before, let Φ be the standard normal cdf.

Theorem 2.2.1. (Berry-Esseen theorem) *Let $\{X_n\}$ be a sequence of IID random variables with $EX_1 = 0$, $E[X_1^2] = \sigma^2$ and $E[|X_1|^3] = \delta_3 < \infty$. Then*

$$\|F_n - \Phi\| \leq 3\delta_3/\sigma^3\sqrt{n} \quad \text{for all } n,$$

where $F_n(x) \equiv P((n\sigma^2)^{-1/2}(X_1 + \cdots + X_n) \leq x)$.

Theorem 2.2.1 implies that for given n and σ^2 , the bound on the distances decreases as the third absolute moment δ_3 decreases. We now describe the Edgeworth expansion, which shows how further regularity conditions can improve the quality of the normal approximation; see p. 535 of Feller (1971). We also get convergence of pdf's.

Theorem 2.2.2. (Edgeworth expansion) *If, in addition to the assumptions of Theorem 2.2.1 above, moments $E[X_1^k]$ exist for $3 \leq k \leq r$ and $|E[\exp(itX_1)]|^\nu$ is integrable for some $\nu \geq 1$, then $(n\sigma^2)^{-1/2}(X_1 + \cdots + X_n)$ has a pdf f_n for all n and*

$$f_n(x) = n(x)[1 + \sum_{k=3}^r n^{-(k-2)/2} P_k(x) + o(n^{-(r-2)/2})]$$

as $n \rightarrow \infty$, uniformly in x , where n is the standard normal pdf and $P_k(x)$ is a real polynomial depending on the first k moments of X_1 , with the property that $P_k(x) = 0$ if the first k moments of X_1 agree with those of the standard normal distribution.

Note that the rate of convergence in Theorem 2.2.2 is $O(n^{-1/2})$ if $E[X_1^3] \neq 0$, but is $O(n^{-1})$ or better if $E[X_1^3] = 0$. When $E[X_1^3] \neq 0$, the refinement provided by the second term can be useful.

2.2.2. Rates of Convergence in the FCLT

We now turn to Donsker's FCLT. From the Lipschitz mapping theorem, Theorem 3.4.2 in the book, we can deduce a bound on the rate of convergence in the CLT from a bound on a rate of convergence in the FCLT. Hence, we can see in advance that the rate of convergence in the FCLT, given a finite third absolute moment, can be no better than the $O(n^{-1/2})$ bound provided by the Berry-Esseen theorem. In fact, the best possible bound for the FCLT, under an even stronger regularity condition, is somewhat worse, being larger by a factor of $\log n$. From a practical perspective, though, the difference is not great.

We now give the final rate-of-convergence result, expressed in terms of the Prohorov metric π from Section 3.2 of the book; see (2.2) here. For this application, it is convenient to let the underlying function space be the set $D_Q \equiv D_Q([0, 1], \mathbb{R})$ of functions in $D \equiv D([0, 1], \mathbb{R})$ with discontinuities only at rational points in the domain $[0, 1]$, endowed with the uniform metric $\|\cdot\|$; we refer to the space as (D_Q, U) . The space (D_Q, U) is a separable metric space and the stochastic processes considered here all have sample paths in this space. Thus, the Prohorov metric π is defined on the space $\mathcal{P}((D_Q, U))$, the space of all probability measures on (D_Q, U) . Since

$$d_{M_1}(x_1, x_2) \leq d_{J_1}(x_1, x_2) \leq \|x_1 - x_2\| \quad \text{for } x_1, x_2 \in D,$$

the result also holds for the spaces (D, d_{J_1}) and (D, d_{M_1}) .

The following combines Theorems 1.16 and 1.17 in Csörgő and Horváth (1993).

Theorem 2.2.3. (bounds on the rate of convergence in Donsker's FCLT) *Let $\{X_n\}$ be a sequence of IID random variables with $EX_n = 0$ and $E[X_1^2] = \sigma^2$. If, in addition, $E[\exp(tX_1)] < \infty$ for t in a neighborhood of the origin, then there exist positive constants C_1 and C_2 such that*

$$C_1 \log n / \sqrt{n} \leq \pi(\mathbf{S}_n, \sigma \mathbf{B}) \leq C_2 \log n / \sqrt{n} \quad (2.2)$$

for all n , where π is the Prohorov metric on the space $\mathcal{P}((D_Q, U))$, $\mathbf{B}$ is standard Brownian motion and $\mathbf{S}_n(t) \equiv n^{-1/2}S_{[nt]}$, $0 \leq t \leq 1$. If, instead, only $E[|X_1|^p] < \infty$ for some $p > 2$, then there is a constant C such that

$$\pi(\mathbf{S}_n, \sigma \mathbf{B}) \leq C n^{-(p-2)/2(p+1)} \quad (2.3)$$

for all n . Moreover, for any sequence $\{a_n\}$ with $a_n \rightarrow \infty$ as $n \rightarrow \infty$, there is a random variable X_1 with $E[|X_i|^p] < \infty$ such that

$$\overline{\lim}_{n \rightarrow \infty} a_n n^{(p-2)/2(p+1)} \pi(\mathbf{S}_n, \sigma \mathbf{B}) = \infty. \quad (2.4)$$

The lower bound in (2.2) and the limit in (2.4) show that the upper bounds in Theorem 2.2.3 are indeed best possible. Note that the rate $O(\log n/\sqrt{n})$ in (2.2) exceeds the Berry-Esseen bound $O(1/\sqrt{n})$ by a factor of $\log n$. We regard that difference as negligible.

However, there is a big difference between the bounds in (2.3) and in Theorem 2.2.2. When there is only a finite third absolute moment, we have (2.3) with $p = 3$, which only yields the rate $O(n^{-1/8})$. For finite p^{th} moment with $p > 2$, (2.3) gives a rate that can be substantially worse than $O(n^{-1/2})$, while Theorem 2.2.2 gives rates that can be much better than $O(n^{-1/2})$. It should be recognized that the conditions are quite different though.

By the Lipschitz mapping theorem, Theorem 3.4.2 of the book, the rate of convergence in Theorem 2.2.3 is inherited by Lipschitz functions. For real-valued Lipschitz functions, we then can obtain bounds on the uniform metric for cdf's.

Corollary 2.2.1. (bounds on the uniform metric for cdf's of the images of real-valued Lipschitz maps) *Suppose that $g : (D_Q, U) \rightarrow \mathbb{R}$ is a Lipschitz function and that $g(\mathbf{B})$ has a bounded pdf. If the conditions of Theorem 2.2.3 hold with $E\exp(tX_1) < \infty$ for t in a neighborhood of the origin, then there is a positive constant C such that*

$$\sup_x |P(g(\mathbf{S}_n) \leq x) - P(g(\sigma\mathbf{B}) \leq x)| \leq C \log n / \sqrt{n} \quad (2.5)$$

for all $n \geq 1$.

We can apply Corollary 2.2.1 to obtain a bound on the rate of convergence in the CLT; we use the projection map $\pi_1(x) \equiv x(1)$, which is easily seen to be Lipschitz. However, the bound is not as good as provided by the Berry-Esseen theorem, so the bound may no longer be best possible when we consider the image measure associated with a single Lipschitz map.

We can also apply Theorem 2.2.3 to establish bounds on the rate of convergence in heavy-traffic FCLTs for queues. We illustrate by stating a result for the queueing model in Section 1.6. We use the fact that the two-sided reflection map $\phi_K : D \rightarrow D$ is Lipschitz; see Theorem 13.10.1. An early result of this kind is Kennedy (1973). That served as motivation for the Lipschitz mapping theorem in Whitt (1974).

Corollary 2.2.2. (bounds on the rate of convergence in a heavy-traffic stochastic-process limit for queues) *Consider the queueing model in Section 2.3 of the book with IID inputs V_k with mean m_v and variance σ^2 .*

2.2. STRONG APPROXIMATIONS AND RATES OF CONVERGENCE 27

If, in addition, $K_n = n^{1/2}K$ and $\mu_n = m_v + mn^{-1/2}$ for all n and with $E[\exp(tV_1)] < \infty$ for some $t > 0$, then there exists a constant C such that

$$\pi(\mathbf{W}_n, \phi_K(\sigma\mathbf{B} - m\mathbf{e})) \leq C \log n / n^{1/2} ,$$

where $\mathbf{W}_n$ is the scaled workload process in equation (2.3.6) of the book and ϕ_K is the two-sided reflection map.

2.2.3. Strong Approximations

Theorem 2.2.3 can be established by applying *strong approximations*. Like the Skorohod and Strassen representation theorems in Chapters 3 and 11 of the book, strong approximations are special constructions of random objects on the same underlying probability space, often called couplings; see Lindvall (1992).

We start by stating the Komlós, Major and Tusnády (1975, 1976) strong approximation theorems for partial sums of IID random variables; see Chapter 2 of Csörgő and Révész (1981) and Chapter 1 of Csörgő and Horváth (1993). See Philipp and Stout (1975) for extensions to the weakly dependent case and Einmahl (1989) for extensions to the multivariate case. See Csörgő and Horvath (1993) for strong approximations of renewal processes and random sums. For applications of strong approximations to queues, see Zhang et al. (1990), Horváth (1990), Glynn and Whitt (1991a,b) and Chen and Mandelbaum

Theorem 2.2.4. (strong approximation with finite moment generating function) *Let $\{X_n : n \geq 1\}$ be a sequence of IID random variables with $EX_1 = 0$, $EX_1^2 = 1$ and $Ee^{tX_1} < \infty$ for t in a neighborhood of the origin. Let $S_n \equiv X_1 + \cdots + X_n$, $n \geq 1$, with $S_0 \equiv 0$. Then there exists a standard Brownian motion $\mathbf{B} \equiv \{\mathbf{B}(t) : t \geq 0\}$ such that, for all real x and every $n \geq 1$,*

$$P \left(\max_{1 \leq k \leq n} |S_k - \mathbf{B}(k)| > C_1 \log n + x \right) < C_2 e^{-\lambda x} , \quad (2.6)$$

where C_1 , C_2 and λ are positive constants depending upon the distribution of X_1 .

As a consequence of Theorem 2.2.4, we can deduce that

$$S_n - \mathbf{B}(n) = O(\log n) \quad \text{as } n \rightarrow \infty \quad \text{w.p.1} ; \quad (2.7)$$

i.e., there is a constant C such that

$$P(|S_n - \mathbf{B}(n)| > C \log n \text{ infinitely often}) = 0 . \quad (2.8)$$

Note that (2.8) follows from (2.6) by substituting $C' \log n$ for x in (2.6) for suitably large C' and then applying the Borel-Cantelli theorem.

We now relax the extra condition on the tail of the cdf $P(|X_1| > t)$, at the expenses of obtaining a slower rate.

Theorem 2.2.5. (strong approximation with p^{th} moment) *Let $\{X_n : n \geq 1\}$ be a sequence of IID random variables with $EX_1 = 0$, $EX_1^2 = 1$ and $E|X_1|^p < \infty$ for some $p > 2$. Let $S_n \equiv X_1 + \cdots + X_n$, $n \geq 1$, with $S_0 \equiv 0$. Then there exists a standard Brownian motion $\mathbf{B}$ such that*

$$n^{-1/p}|S_n - \mathbf{B}(n)| \rightarrow 0 \quad \text{w.p.1} \quad (2.9)$$

To apply Theorems 2.2.4 and 2.2.5 to establish Theorem 2.2.3, we need to relate Brownian motion $\mathbf{B}$ to the associated processes

$$\mathbf{B}_n^1(t) \equiv n^{-1/2}\mathbf{B}(\lfloor nt \rfloor), \quad \mathbf{B}_n^2(t) \equiv \mathbf{B}(\lfloor nt \rfloor/n), \quad \mathbf{B}_n^3(t) \equiv n^{-1/2}\mathbf{B}(nt)$$

for $0 \leq t \leq 1$. By the self-similarity property, $\mathbf{B} \stackrel{d}{=} \mathbf{B}_n^3$ and $\mathbf{B}_n^1 \stackrel{d}{=} \mathbf{B}_n^2$ for all $n \geq 1$. We can relate $\mathbf{B}_n^2$ to $\mathbf{B}$ by bounding the fluctuations of Brownian motion. The following is Lemma 1.1.1 of Csörgő and Révész (1981).

Theorem 2.2.6. (uniform bound on the fluctuations of Brownian motion) *For any $\epsilon > 0$, there exists a constant $C = C(\epsilon)$ such that*

$$P\left(\sup_{0 \leq t \leq T-h} \sup_{0 \leq s \leq h} |\mathbf{B}(t+s) - \mathbf{B}(t)| \geq \nu\sqrt{h}\right) \leq (CT/h)\exp(-\nu^2/(2+\epsilon)) \quad (2.10)$$

for all positive ν , T , and h , $0 < h < T$.

Theorem 2.2.6 can be applied to determine the precise modulus of continuity of Brownian sample paths (originally determined by Lévy); see Theorem 1.1 of Csörgő and Révész (1981).

Theorem 2.2.7. (modulus of continuity of Brownian paths) *If $\mathbf{B}$ is Brownian motion, then*

$$\lim_{h \rightarrow 0} \sup_{0 \leq s \leq 1} \sup_{0 \leq t \leq h} \frac{|\mathbf{B}(s+t) - \mathbf{B}(s)|}{\sqrt{2h \log h^{-1}}} = 1 \quad \text{w.p.1}.$$

From Theorem 2.2.7, we see that the sample paths of Brownian motion are continuous but not differentiable; the largest increment of length h is almost surely of order $O(\sqrt{2h \log h^{-1}})$. We can also apply Theorem 2.2.6 to determine the following bound on the in-probability distance $p(\mathbf{B}, \mathbf{B}_n^2)$ and the Prohorov distance $\pi(\mathbf{B}, \mathbf{B}_n^1)$, where π is defined on the space $\mathcal{P}((C, U))$.

Corollary 2.2.3. *There exists a constant C_1 such that*

$$\pi(\mathbf{B}, \mathbf{B}_n^1) \leq p(\mathbf{B}, \mathbf{B}_n^2) \leq C_1 \sqrt{\log n/n}$$

for all $n \geq 1$.

Proof. The first inequality holds because $\mathbf{B}_n^1 \stackrel{d}{=} \mathbf{B}_n^2$ and $\pi \leq p$. For the second inequality, let $\nu = \sqrt{c \log n}$ for $c > 4$ in (2.10). Then the right hand side of (2.10) for $T = 1$ becomes $C'n^{-(1+\delta)}$ for $\delta > 0$ and constant C' . ■

Partial proof of Theorem 2.2.3. For the upper bound in (2.1), let $x = C_3 \log n$ in (2.6) to obtain

$$\pi(\mathbf{S}_n, \mathbf{B}_n^1) \leq p(\mathbf{S}_n, \mathbf{B}_n^1) \leq C \log n / \sqrt{n}.$$

Then use the triangle inequality with Corollary 2.2.3. ■

Theorem 2.2.4 can be applied to obtain a strong approximation for a Lévy process, i.e., a random element of D with stationary and independent increments; see Corollary 5.5 on p. 359 of Ethier and Kurtz (1986).

Theorem 2.2.8. (strong approximation for a Lévy process) *Let $\{\mathbf{L}(t) : t \geq 0\}$ be a real-valued Lévy process. Assume that*

$$Ee^{\alpha \mathbf{L}(1)} < \infty \tag{2.11}$$

for all α with $|\alpha| \leq \alpha_0$ for some $\alpha_0 > 0$. Then there exist versions of the Lévy process $\mathbf{L}$ and a standard Brownian motion $\mathbf{B}$ on a common probability space such that

$$|\mathbf{L}(t) - mt - \sigma \mathbf{B}(t)| = O(\log t) \quad \text{as } t \rightarrow \infty \quad \text{w.p.1}, \tag{2.12}$$

where $m = E\mathbf{L}(1)$ and $\sigma^2 = \text{Var } \mathbf{L}(1)$.

A precursor to the strong approximation theorems, of interest in its own right, is the Skorohod (1961) embedding theorem; see p. 88 of Csörgő and Révész (1981).

Theorem 2.2.9. (Skorohod embedding theorem) *Let $\{X_n : n \geq 1\}$ be a sequence of IID real-valued random variables with $EX_1 = 0$ and $EX_1^2 = 1$. Let $S_n \equiv X_1 + \cdots + X_n$, $n \geq 1$, with $S_0 \equiv 0$. There exists a probability space supporting a standard Brownian motion $\mathbf{B}$ and a sequence $\{T_n : n \geq 1\}$ of nonnegative IID random variables such that*

- (i) $\{\mathbf{B}(T_1 + \cdots + T_n) : n \geq 1\} \stackrel{d}{=} \{S_n : n \geq 1\}$ in $\mathbb{R}^\infty$;
- (ii) $\{T_1 + \cdots + T_n : n \geq 1\}$ is a sequence of stopping times, i.e., the event $\{T_1 + \cdots + T_n \leq t\}$ is contained in the σ -field generated by $\{\mathbf{B}(s) : 0 \leq s \leq t\}$ for all $t \geq 0$;
- (iii) $ET_1 = 1$;
- (iv) $ET_1^k < \infty$ if, in addition, $EX^{2k} < \infty$ for positive integer k .

As a consequence of Theorem 2.2.9,

$$\begin{aligned} \{n^{-1/2}S_{[nt]} : t \geq 0\} &\stackrel{d}{=} \{n^{-1/2}B(T_1 + \cdots + T_{[nt]}) : t \geq 0\} \\ &\stackrel{d}{=} \{B(n^{-1}(T_1 + \cdots + T_{[nt]})) : t \geq 0\} . \end{aligned}$$

By the FSLN,

$$\sup_{0 \leq t \leq u} |n^{-1}(T_1 + \cdots + T_{[nt]}) - t| \rightarrow 0 \quad \text{w.p.1} ,$$

so that Donsker's theorem again is a consequence. Rate of convergence results follow too.

2.3. Weak Dependence from Regenerative Structure

This section is a sequel to Section 4.4 in the book, in which we showed that many Brownian limits still hold for random walks $\{S_n : n \geq 0\}$ when the IID condition on the sequence of steps $\{X_n : n \geq 1\}$ is relaxed, with the finite-second-moment condition $EX_n^2 < \infty$ remaining in place. We now obtain results for stochastic-processes with regenerative structure.

This new setting allows us to abandon the assumption of stationarity and obtain explicit expressions for the asymptotic variance σ^2 , defined by

$$\sigma^2 \equiv \lim_{n \rightarrow \infty} \frac{\text{Var}(S_n)}{n} . \quad (3.1)$$

For a stationary sequence $\{X_n\}$, the asymptotic variance has the representation

$$\sigma^2 = \text{Var } X_n + 2 \sum_{k=1}^{\infty} \text{Cov}(X_1, X_{1+k}) . \quad (3.2)$$

We now obtain more explicit representations for the asymptotic variance in terms of basic model elements.

2.3.1. Discrete-Time Markov Chains

We start by stating results for finite-state Markov chains. We first consider discrete-time chains and then we consider continuous-time chains. Afterwards, we state results for general regenerative processes, which cover more general Markov processes and non-Markov processes. The first result for DTMC's extends Theorem 4.4.2 in the book. An important point is that an explicit expression can be given for the asymptotic variance σ^2 . It is expressible as a function of the fundamental matrix of the DTMC. The most effective way to calculate the asymptotic variance is usually to solve a system of equations, collectively known as the *Poisson equation*.

Let P be the transition matrix of an irreducible k -state DTMC and let Π be a matrix with each row being the steady-state vector π . (We will work with row vectors; let A^t be the transpose of a matrix A , so that the column vector associated with a row vector x is x^t .) Then the *fundamental matrix* of the DTMC is

$$Z \equiv (I - P + \Pi)^{-1} ; \quad (3.3)$$

see pp. 75, 100 of Kemeny and Snell (1960). (The matrix $I - P + \Pi$ is nonsingular.)

Theorem 2.3.1. (FCLT for a DTMC with explicit asymptotic variance)
Let $\{Y_n : n \geq 1\}$ be an irreducible k -state DTMC and let $X_n = f(Y_n)$ for a real-valued function f . Then the FCLT

$$\mathbf{S}_n \Rightarrow \sigma \mathbf{B} \quad \text{in } (D, J_1) , \quad (3.4)$$

where $\mathbf{B}$ is standard Brownian motion and

$$\mathbf{S}_n(t) = n^{-1/2}(S_{[nt]} - mnt), \quad t \geq 0 , \quad (3.5)$$

holds with

$$m \equiv \sum_{i=1}^k \pi_i f(i) ,$$

$$\sigma^2 \equiv 2 \sum_{i=1}^k \sum_{j=1}^k (f(i) - m) \pi_i Z_{i,j} (f(j) - m) - \sum_{i=1}^k \pi_i (f(i) - m)^2 , \quad (3.6)$$

π the steady-state vector and $Z \equiv (Z_{i,j})$ the fundamental matrix in (3.3).

As a quick sanity check on (3.6), note that in the IID case we have $P = A$, $Z = I$ and, from (3.6),

$$\sigma^2 = \sum_{i=1}^k \pi_i (f(i) - m)^2 ,$$

as we should.

It is significant that we can calculate π , m , Z and σ^2 in Theorem 2.3.1 by solving the Poisson equation(s). We state both row-vector and column-vector versions. Let $\mathbf{1} \equiv (1, \dots, 1)$ be a vector of 1's and $\mathbf{0} \equiv (0, \dots, 0)$ be a vector of 0's.

Theorem 2.3.2. (Poisson equations for a DTMC) *Consider an irreducible finite-state DTMC with transition matrix P . The row-vector version of the Poisson equation*

$$x(I - P) = y \tag{3.7}$$

has a solution x for given y if and only if $y\mathbf{1}^t = 0$. All solutions to (3.7) are of the form

$$x = yZ + (x\mathbf{1}^t)\pi .$$

The column-vector version of the Poisson equation

$$(I - P)x^t = y^t \tag{3.8}$$

has a solution x^t for given y^t if and only if $\pi y^t = 0$. All solutions to (3.8) are of the form

$$x^t = Zy^t + (\pi x^t)\mathbf{1} .$$

Proof. We consider only the row-vector form. Clearly $y\mathbf{1}^t = 0$ is necessary, because $(I - P)\mathbf{1}^t = \mathbf{0}^t$. Given (3.7),

$$x(I - P + \Pi) = y + (x\mathbf{1}^t)\pi ,$$

but $I - P + \Pi$ is nonsingular with inverse Z , so that

$$x = yZ + (x\mathbf{1}^t)\pi Z = yZ + (x\mathbf{1}^t)\pi$$

since $\pi Z = \pi$. ■

Theorem 2.3.3. (Poisson equations for the steady-state vector and the asymptotic variance of a DTMC) *For an irreducible finite-state DTMC, the steady-state vector π is the unique solution x to the Poisson equation (3.7)*

2.3. WEAK DEPENDENCE FROM REGENERATIVE STRUCTURE 33

with $y = (0, \dots, 0)$ and $x\mathbf{1}^t = 1$. The asymptotic variance can be expressed as

$$\sigma^2 = 2 \sum_{i=1}^k x_i (f(i) - m)$$

where m is the mean and x solves the Poisson equation (3.7) with

$$y_i = (f(i) - m)\pi_i, \quad 1 \leq i \leq k .$$

2.3.2. Continuous-Time Markov Chains

We now turn to the continuous-time processes. There are analogs of the DTMC results in Theorems 2.3.1–2.3.3 for CTMC's. Let $\{(Y(t) : t \geq 0\}$ be an irreducible k -state CTMC. Then the limit is for the integral

$$S(t) \equiv \int_0^t f(Y(s)) ds, \quad t \geq 0 .$$

The associated normalized processes in D are

$$\mathbf{S}_n(t) \equiv n^{-1/2}(S(nt) - mnt), \quad t \geq 0 . \quad (3.9)$$

Given transition matrices $P(t) \equiv (P_{i,j}(t))$, where

$$P_{i,j}(t) \equiv P(Y(t) = j | Y(0) = i) ,$$

the *infinitesimal generator matrix* of the CTMC is $Q \equiv (Q_{i,j})$ where

$$Q \equiv \lim_{t \downarrow 0} (P(t) - I)$$

and the *fundamental matrix* is $Z \equiv (Z_{i,j})$ where

$$Z_{i,j} \equiv \int_0^\infty (P_{i,j}(t) - \pi_j) dt$$

and

$$Z = (\Pi - Q)^{-1} - \Pi \quad (3.10)$$

see Kemeny and Snell (1961) and Whitt (1992). A CTMC model is usually specified by giving the infinitesimal generator matrix Q . For an irreducible finite-state CTMC, the steady-state vector π is the unique vector with sum 1 that satisfies

$$\pi Q = 0 .$$

Paralleling (3.1) and (3.2) above, the asymptotic variance in this continuous-time framework is

$$\sigma^2 \equiv \lim_{t \rightarrow \infty} \frac{\text{Var}(S(t))}{t} = 2 \int_0^\infty r(t) dt ,$$

where $r(t)$ is the (auto) covariance function, i.e.,

$$r(t) \equiv E[X(0)X(t)] - (E[X(0)])^2$$

for $X(t) \equiv f(Y(t))$, $t \geq 0$.

The following is the continuous-time analog of Theorem 2.3.1.

Theorem 2.3.4. (FCLT for a CTMC with explicit asymptotic variance) *Let $\{Y(t) : t \geq 0\}$ be an irreducible k -state CTMC, and let $X(t) = f(Y(t))$ for a real-valued function f . Then the FCLT (3.4) holds for $\mathbf{S}_n$ in (3.9) with m the steady-state mean and σ^2 the asymptotic variance, which can be expressed as*

$$\sigma^2 \equiv 2 \sum_{i=1}^k \sum_{j=1}^k f(i) \pi_i Z_{i,j} f(j) ,$$

where Z is the fundamental matrix in (3.10).

We can calculate π , m , Z and σ^2 by solving Poisson equations for CTMC's; see Whitt (1992). The following is the continuous-time analog of Theorem 2.3.2.

Theorem 2.3.5. (Poisson equations for a CTMC) *Consider an irreducible finite-state CTMC with infinitesimal generator matrix Q . The row-vector version of the Poisson equation*

$$xQ = y \tag{3.11}$$

has a solution x for given y if and only if $y\mathbf{1}^t = 0$. All solutions to (3.11) are of the form

$$x = -yZ + (x\mathbf{1}^t)\pi .$$

The column-vector version of the Poisson equation

$$Qx^t = y^t$$

has a solution x^t for given y^t if and only if $\pi y^t = 0$. All solutions are of the form

$$x^t = -Zy^t + (\pi x^t)\mathbf{1}^t .$$

2.3. WEAK DEPENDENCE FROM REGENERATIVE STRUCTURE 35

The following is the continuous-time analog of Theorem 2.3.3.

Theorem 2.3.6. (Poisson equations for the steady-state vector and the asymptotic variance of a CTMC) *For an irreducible finite-state CTMC, the steady-state vector π is the unique solution x to the Poisson equation (3.11) with $y = (0, \dots, 0)$ and $x\mathbf{1}^t = 1$. The asymptotic variance can be expressed as*

$$\sigma^2 = 2 \sum_{i=1}^k x_i f_i ,$$

where x is the unique solution to the Poisson equation (3.11) with

$$y_i = (f_i - m)\pi_i \quad \text{and} \quad \sum_{i=1}^k x_i = 0 .$$

and m is the mean.

We can also obtain even more explicit expressions for the asymptotic variance in Markov chains with additional structure. For example, suppose that the CTMC $\{Y(t) : t \geq 0\}$ is a birth-and-death processes on the integers $\{0, 1, \dots, n\}$ with positive birth rates λ_i , death rates μ_i and stationary probabilities

$$\pi_i = \frac{\pi_0 \lambda_0 \lambda_1 \cdots \lambda_{i-1}}{\mu_1 \mu_2 \cdots \mu_i} . \quad (3.12)$$

If the process is irreducible, then the process must be reflecting at 0 and n ; i.e., $\lambda_n = \mu_0 = 0$.) The following is Proposition 1 of Whitt (1992). Corresponding results for diffusion processes are also stated there.

Theorem 2.3.7. (asymptotic variance of a birth-and-death process) *Suppose that $X(t) = f(Y(t))$, where f is a real-valued function and $\{Y(t) : t \geq 0\}$ is an irreducible birth-and-death process on the integers $\{0, 1, \dots, n\}$ with birth rates λ_i and death rates μ_i . Then the asymptotic variance can be expressed as*

$$\sigma^2 = 2 \sum_{j=0}^{n-1} (\lambda_j \pi_j)^{-1} \left[\sum_{i=0}^j (f(i) - m) \pi_i \right]^2$$

for m the mean and π in (3.12) above.

We now state a corollary of Theorem 2.3.7 for an elementary queueing model – the $M/M/1$ queue. The queue-length process in an $M/M/1$ queue

is a birth-and-death process with $\lambda_i = \lambda$ and $\mu_i = \mu$ when positive. The following would properly be a corollary to Theorem 2.3.7 except for the fact that the state space is infinite. Extensions to countably infinite and more general state spaces are covered by the results for regenerative processes below.

Corollary 2.3.1. (asymptotic variance for the queue-length process in the M/M/1 queue) *For the queue-length (number in system) process in the M/M/1 queue with traffic intensity $\rho \equiv \lambda/\mu < 1$, the asymptotic variance is*

$$\sigma^2 = \frac{2\rho(1+\rho)}{(1-\rho)^4}. \quad (3.13)$$

The $(1-\rho)^4$ term in the denominator of (3.13) shows that very long simulation runs are required to directly estimate the steady-state mean of the queue-length process by the sample mean when ρ is close to its upper limit 1. That insight is important for related models for which we do not already know the steady-state distribution, so that simulation is actually needed. We discuss applications of stochastic-process limits to obtain insights about simulation in Section 5.9 of the book.

For a birth-and-death process it is also possible, and usually preferable, to recursively solve the Poisson equation, see Remarks 1, 2 and 5 of Whitt (1992). For more on Poisson equations, see Glynn (1994) and Glynn and Meyn (1996).

2.3.3. Regenerative FCLT

Donsker's theorem itself applies quite directly when we have regenerative structure, as in the case of DTMC's and CTMC's in Theorem 2.3.1 and 2.3.4 above. For this discussion, we use the classical definition of regenerative process, meaning that the process splits into IID cycles; see p. 125 of Asmussen (1987). We will present the result in continuous time, following Glynn and Whitt (1993), but corresponding results hold in discrete time, as in Glynn and Whitt (1987). An earlier related Markov chain FCLT is due to Maigret (1978).

Consider a stochastic process $\{Y(t) : t \geq 0\}$ with general state space and a measurable real-valued function f on that state space. We assume that the stochastic process $\{Y(t) : t \geq 0\}$ is regenerative with respect to regeneration times T_i satisfying

$$0 \leq T_0 < T_1 < \dots$$

2.3. WEAK DEPENDENCE FROM REGENERATIVE STRUCTURE 37

with $T_{-1} \equiv 0$. We focus on the associated *cumulative process*

$$C(t) \equiv \int_0^t f(Y(s))ds, \quad t \geq 0, \quad (3.14)$$

and consider the associated normalized processes

$$\mathbf{C}_n(t) \equiv n^{-1/2}(C(nt) - mnt), \quad t \geq 0 \quad (3.15)$$

where m is a real number yet to be specified. The key random variables associated with the regenerative cycles are

$$\begin{aligned} \tau_i &\equiv T_i - T_{i-1}, \\ X_i &\equiv X_i(m) \equiv \int_{T_{i-1}}^{T_i} [f(Y(u)) - m]du, \\ Z_i &\equiv Z_i(m) \equiv \sup_{0 \leq s \leq \tau_i} \left| \int_0^s [f(Y(T_{i-1} + u)) - m]du \right|. \end{aligned} \quad (3.16)$$

By *regenerative structure* we mean that the three-tuples (τ_i, X_i, Z_i) are IID for $i \geq 1$. We also assume that $E\tau_i < \infty$ and

$$\int_0^t |f(Y(s))|ds < \infty \text{ w.p.1 for each } t,$$

which implies that the cumulative process has continuous sample paths w.p.1.

The general idea is that the cumulative process C in (3.14) is approximately equal to a random sum. In particular,

$$C(t) = S_{N(t)} + R_1(t) + R_2(t), \quad t \geq 0,$$

where

$$S_n \equiv X_1 + \cdots + X_n, \quad n \geq 1,$$

for X_i in (3.16) with $S_0 \equiv 0$, $N \equiv \{N(t) : t \geq 0\}$ is the (possibly delayed) renewal counting process associated with the regeneration times, i.e.,

$$N(t) \equiv \max\{i : T_i \leq t\}, \quad t \geq 0,$$

and $R_i \equiv \{R_i(t) : t \geq 0\}$ are remainder processes, defined by

$$R_1(t) \equiv \int_0^{\min\{t, T_0\}} f(Y(s))ds \quad (3.17)$$

and

$$R_2(t) = \int_{T_{N(t)}}^t f(Y(s))ds, \quad t \geq 0. \quad (3.18)$$

Since $E\tau_1 < \infty$, we have

$$t^{-1}N(t) \rightarrow \lambda \equiv 1/E\tau_1, \quad \text{as } t \rightarrow \infty \quad \text{w.p.1} . \quad (3.19)$$

Under (3.19), FCLTs for partial sums tend to extend to random sums, as we see in Chapter 13 of the book. The major difficulty here is treating the two remainder terms in (3.17). Since $|R_1(t)| \leq Z_0$, the first remainder term in (3.17) is easily dispensed with in limit theorems. The second remainder term is more complicated; the key bound is

$$|R_2(t)| \leq Z_{N(t)+1}, \quad t \geq 0 .$$

Then we observe that $\{R_2(t) : t \geq 0\}$ is tight without space scaling. Thus, after space scaling, it is asymptotically negligible.

Theorem 2.3.8. (FCLT for regenerative processes) *With the regenerative structure above, there is convergence in distribution*

$$\mathbf{C}_n \Rightarrow \sigma \mathbf{B} \quad \text{in } (D, J_1)$$

for $\mathbf{C}_n$ in (3.15) and $\mathbf{B}$ standard BM if and only if there is a constant m such that

$$EX_1(m) = 0, \quad EX_1(m)^2 < \infty$$

and

$$t^2 P(Z_1(m) > t) \rightarrow 0 \quad \text{as } t \rightarrow \infty . \quad (3.20)$$

for $X_1(m)$ and $Z_1(m)$ in (3.16). Then the asymptotic variance is

$$\sigma^2 = EX_1(m)^2 .$$

A sufficient condition for the regularity condition (3.20) is $EZ_1(m)^{2+\epsilon} < \infty$ for some $\epsilon > 0$. (A finite second moment is not enough. We remark that condition (3.20) does not appear in the ordinary CLT; see Glynn and Whitt (1993, 2000).) The role of the regularity condition (3.20) can be understood from the following lemma.

Lemma 2.3.1. (condition for the scaled maximum to be asymptotically negligible) *Let $\{Z_i : i \geq 1\}$ be a sequence of IID real-valued random variables and let $\psi : R_+ \rightarrow R_+$ be a function such that $\psi(t) \rightarrow \infty$ as $t \rightarrow \infty$. Then*

$$\psi(n)^{-1} \max_{1 \leq i \leq n} \{Z_i\} \Rightarrow 0 \quad \text{as } n \rightarrow \infty$$

2.3. WEAK DEPENDENCE FROM REGENERATIVE STRUCTURE 39

if and only if

$$tP(|Z_1| > \epsilon\psi(t)) \rightarrow 0 \quad \text{as } t \rightarrow \infty \quad \text{for all } \epsilon > 0. \quad (3.21)$$

Proof. Let $M_n \equiv \max\{|Z_i| : 1 \leq i \leq n\}$ and $F(t) \equiv P(|Z_1| \leq t)$, $t \geq 0$. Note that $\psi(n)^{-1}M_n \Rightarrow 0$ if and only if, for all $\epsilon > 0$, $P(\psi(n)^{-1}M_n > \epsilon) \rightarrow 0$ as $n \rightarrow \infty$. However,

$$P(M_n > \epsilon\psi(n)) < \delta$$

if and only if

$$P(M_n \leq \epsilon\psi(n)) \geq 1 - \delta,$$

where

$$\begin{aligned} P(M_n \leq \epsilon\psi(n)) &= F(\epsilon\psi(n))^n \\ &= (1 - n^{-1}n(1 - F(\epsilon\psi(n))))^n \\ &= (1 - n^{-1}nF^c(\epsilon\psi(n)))^n \\ &\rightarrow 1 \quad \text{as } n \rightarrow \infty \end{aligned} \quad (3.22)$$

if and only if

$$nF^c(\epsilon\psi(n)) \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

or, equivalently, (3.21). ■

Corollary 2.3.2. *If the conditions of Lemma 2.3.1 hold with $\psi(t) = t^\alpha$ for $\alpha > 0$, then condition (3.21) is equivalent to*

$$t^{1/\alpha}P(|Z_1| > t) \rightarrow 0 \quad \text{as } t \rightarrow \infty.$$

Proof. Under the assumption, condition (3.21) becomes

$$tP(|Z_1| > \epsilon t^\alpha) \rightarrow 0 \quad \text{as } t \rightarrow \infty \quad \text{for all } \epsilon > 0,$$

which first is equivalent to

$$\epsilon^\alpha(\epsilon^{-\alpha}t)P(|Z_1| > (\epsilon^{-\alpha}t)^\alpha)$$

and then is equivalent to

$$\epsilon^\alpha tP(|Z_1| > t^\alpha) \rightarrow 0 \quad \text{as } t \rightarrow \infty \quad \text{for all } \epsilon > 0,$$

which in turn is equivalent to the stated result. ■

A general application of Theorem 2.3.8 is to obtain a FCLT for the counting processes associated with a batch Markovian arrival process (BMAP) as in Lucantoni (1993) or, equivalently, the virtual Markovian point process in Neuts (1989). An explicit formula for the variance of the number of arrivals in $[0, t]$ in a BMAP, from which the asymptotic variance easily can be obtained, is given on p. 284 of Neuts (1989).

2.3.4. Martingale FCLT

Martingale FCLTs are versatile tools for many applications. We have stated one martingale FCLT in Theorem 4.4.4 of the book, but there are others. We conclude this section by stating another. It is Theorem 18.1 of Billingsley (1999).

We start with the double sequence $\{X_{n,i} : n \geq 1, i \geq 1\}$ and an associated double sequence of σ -fields $\{\mathcal{F}_{n,k} : n \geq 1, k \geq 1\}$. We assume that $X_{n,k}$ is a *martingale difference* with respect to these σ -fields, i.e., $X_{n,k}$ is $\mathcal{F}_{n,k}$ -measurable and

$$E[X_{n,k} | \mathcal{F}_{n,k-1}] = 0 \quad \text{for all } n \text{ and } k.$$

Suppose that $EX_{n,k}^2 < \infty$ and put

$$V_{n,k} \equiv E[X_{n,k}^2 | \mathcal{F}_{n,k-1}]. \quad (3.23)$$

Note that $V_{n,k}$, being a conditional expectation, is a random variable. If the martingale is originally defined only for $1 \leq k \leq k_n$, let $X_{n,k} = 0$ and $\mathcal{F}_{n,k} = \mathcal{F}_{n,k_n}$ for $k > k_n$. Assume that $\sum_{k=1}^{\infty} X_{n,k}$ and $\sum_{k=1}^{\infty} V_{n,k}$ converge w.p.1 for each n .

Theorem 2.3.9. (martingale FCLT) *If, in addition to the assumptions above,*

$$\sum_{k=1}^{\lfloor nt \rfloor} V_{n,k} \Rightarrow \sigma^2 t \quad \text{as } n \rightarrow \infty \quad \text{for every } t > 0 \quad (3.24)$$

with $V_{n,k}$ in (3.23) and the Lindeberg condition

$$\sum_{k=1}^{\lfloor nt \rfloor} E[X_{n,k}^2 I_{\{|X_{n,k}| \geq \epsilon\}}] \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

holds for every $t > 0$ and $\epsilon > 0$, then

$$\mathbf{S}_n \Rightarrow \sigma \mathbf{B} \quad \text{in } D,$$

where σ is determined by (3.24),

$$\mathbf{S}_n(t) = \sum_{k=1}^{\lfloor nt \rfloor} X_{n,k}, \quad t \geq 0,$$

and $\mathbf{B}$ is standard Brownian motion.

Generalizations and other variations of Theorem 2.3.9 are contained on p. 339 of Ethier and Kurtz (1986) and Jacod and Shiryaev (1987).

2.4. Double Sequences and Lévy Limits

We have seen that there are only a few possible limits for normalized partial-sum processes with weak dependence when we work in the framework of a single sequence $\{X_n : n \geq 1\}$. In addition to the Brownian motion limits discussed in Sections 4.3 and 4.4 of the book, there are the stable Lévy motion limits discussed in Sections 4.5 and 4.7 of the book. However, there are many more possible limits for normalized partial-sum processes with weak dependence when we work in the framework of a double sequence $\{X_{n,k} : n \geq 1, k \geq 1\}$. We give a brief account in this section.

Throughout this section we assume that the sequence $\{X_{n,k} : k \geq 1\}$ is IID for each n , so that we are in a classic well-studied setting; e.g., see Gnedenko and Kolmogorov (1968) and Feller (1971). Since there is a different sequence for each n , we can incorporate multiplicative and additive normalization constants directly in the variables $X_{n,k}$. Hence we focus on the partial sums

$$S_{n,n} \equiv X_{n,1} + \cdots + X_{n,n} \quad (4.1)$$

without further normalization and the associated random functions in D defined by

$$\mathbf{S}_n(t) \equiv S_{n, \lfloor nt \rfloor}, \quad t \geq 0. \quad (4.2)$$

The class of limits processes in FCLTs for $\mathbf{S}_n$ now are all Lévy processes. As indicated in Section 4.5 of the book, a *Lévy process* $\mathbf{L} \equiv \{\mathbf{L}(t) : t \geq 0\}$ is a stochastic process with sample paths in $D \equiv D([0, \infty), \mathbb{R})$, $\mathbf{L}(0) = 0$ and stationary and independent increments. Brownian motion and stable Lévy motion are important examples of Lévy processes, but there are many more; see Bertoin (1996) and Jacod and Shiryaev (1987).

The distribution of $\mathbf{L}(t)$ for any t is an infinitely divisible distribution. A probability distribution is *infinitely divisible* if for each n it is the n -fold convolution of another probability distribution; i.e., a random variable X has an infinitely divisible distribution if, for all n , there are IID random variables $X_1, \dots, X_n$ (depending upon X and n) such that

$$X \stackrel{d}{=} X_1 + \cdots + X_n.$$

Lévy processes and infinitely divisible distributions are characterized by their characteristic functions. In particular, the one-dimensional marginal distribution of every Lévy process has characteristic function

$$E e^{i\theta L(t)} = e^{t\psi(\theta)},$$

where the *Lévy exponent* $\psi(\theta)$ can be expressed as

$$\psi(\theta) = ib\theta - \frac{\sigma^2\theta^2}{2} + \int_{-\infty}^{\infty} (\exp(i\theta x) - 1 - i\theta h(x))\mu(dx) , \quad (4.3)$$

with b being the *centering coefficient*, $\sigma^2 \geq 0$ is the *Gaussian coefficient*, μ the *Lévy measure* and h a *truncation function*. There is quite a lot of freedom in the choice of the truncation function h . Following Jacod and Shiryaev (1987, pp. 75) we assume that the truncation function has compact support, is bounded and coincides with x in a neighborhood of the origin. To characterize convergence, we also want h to be continuous. A truncation function with all these properties is

$$h(x) = \begin{cases} x, & 0 \leq x \leq 1 \\ 2 - x, & 1 \leq x \leq 2 \\ -x, & -1 \leq x \leq 0 \\ 2 + x, & -2 \leq x \leq -1 \\ 0, & |x| \geq 2. \end{cases} \quad (4.4)$$

Other truncation functions are considered in the literature. Changing the truncation function h typically changes the centering coefficient b , but does not change the Gaussian coefficient σ^2 or the Lévy measure μ . The Lévy measure has support on $\mathbb{R} - \{0\}$; it is a bonafide measure with

$$\int_{-\infty}^{\infty} \min\{1, x^2\}\mu(dx) < \infty . \quad (4.5)$$

Given a specific truncation function, such as h in (4.4), there is a one-to-one correspondence between Lévy processes, infinitely distributions and the triple of characteristics (b, σ^2, μ) appearing in (4.3), with $\sigma^2 \geq 0$ and μ being a measure on $\mathbb{R} - \{0\}$ satisfying (4.5).

Brownian motion is the special Lévy process with null Lévy measure, i.e., $\mu(A) = 0$ for all measurable subsets A . Non-Gaussian stable Lévy motions with index α are the special cases with $\sigma^2 = 0$ and

$$\mu(dx) = \begin{cases} c^+ x^{-(1+\alpha)}, & x > 0 , \\ c^- |x|^{-(1+\alpha)}, & x < 0 , \end{cases} \quad (4.6)$$

for nonnegative constants c^+ and c^- , where $c^+ + c^- > 0$. From (4.6), we see that the power-tail structure of a stable law is manifested very strongly in the Lévy measure. While the stable law $S_\alpha(\sigma, \beta, \mu)$ has the power-tail

asymptotics in equations 4.5.12 and 4.5.13 in the book, the corresponding Lévy measure has simple power densities on $(0, \infty)$ and $(-\infty, 0)$. A stable Lévy motion is totally skewed to the right, so that $\beta = 1$, (left, so that $\beta = -1$) if and only if $c^- = 0$ ($c^+ = 0$).

The Lévy measure μ characterizes the possible jumps of the Lévy process. Indeed, the jump process of the Lévy process is a Poisson random measure on $\mathbb{R} \times \mathbb{R}^+$ with intensity $\mu(dx)dt$; i.e., the number of jumps in the Lévy process falling in any spatial subinterval $[a, b]$ during time subinterval $[c, d]$ for $a < b$ and $0 < c < d$ has a Poisson distribution with mean $\mu([a, b])|d - c|$. As a simple consequence, if the Lévy measure μ has support in $\mathbb{R}^+$, then the Lévy process has no negative jumps. Thus we know that the totally skewed stable Lévy motion with $\beta = 1$ (and thus $c^- = 0$ in (4.6)) has sample paths without negative jumps.

A complication with Lévy processes is the large (in general, infinite) number of very small jumps. For any $c > 0$, a Lévy process has only finitely many jumps of at least size c in any finite interval w.p.1. However, for any $c > 0$, it can have infinitely many jumps of absolute size less than or equal to c in any finite interval. This large number of small jumps is compensated for by deterministic drift built into the final integral in (4.3), in particular, this drift occurs in the region that the truncation function h is positive. Thus the true process drift is the sum of the drift b and the drift associated with h . In general, the total drift may be infinite, which explains why the representation (4.3) does not separate out all the drift.

It is possible to decompose a Lévy process into the independent sum of component Lévy processes by decomposing the exponent $\psi(\theta)$ in (4.3) into separate pieces; see Theorem 1 of p. 13 of Bertoin (1996) and its proof. The first component Lévy process L_1 has Lévy exponent

$$\psi_1(\theta) \equiv ib\theta - \frac{\sigma^2\theta^2}{2}$$

and is Brownian motion with drift coefficient b and diffusion coefficient σ^2 . The second component Lévy process L_2 has exponent

$$\psi_2(\theta) = \int_{|x| \geq 2} (\exp(i\theta x) - 1)\mu(dx)$$

and is a compound Poisson process, with jumps of absolute size at least 2, having Poisson intensity $\lambda_2 \equiv \mu((-\infty, -2]) + \mu((2, \infty)) < \infty$ and jump size probability distribution $\mu(dx)/\lambda_2$ on $(-\infty, -2) \cup (2, \infty)$. The complicated component is the third one. The third component Lévy process L_3 has

exponent

$$\psi_3(\theta) = \int_{-2}^2 (\exp(i\theta x) - 1 - i\theta h(x))\mu(dx) .$$

It can be shown to be a pure jump martingale with jumps of absolute size at most 2. It includes some deterministic drift to compensate for the jumps. In summary, we can write

$$\psi(\theta) = \psi_1(\theta) + \psi_2(\theta) + \psi_3(\theta)$$

and

$$L \stackrel{d}{=} L_1 + L_2 + L_3 ,$$

where L_1 , L_2 and L_3 are the independent Lévy processes with exponents ψ_1 , ψ_2 and ψ_3 defined above.

If an infinitely divisible distribution has finite moments, these moments can be derived by differentiating the characteristic function. For example, if $E|L(1)| < \infty$, then

$$EL(1) = \frac{\psi'(\theta)}{i} = b + \int_{-\infty}^{\infty} [x - h(x)]\mu(dx) , \quad (4.7)$$

where, because of the definition of the truncation function h , the integrand is non-zero only in $(-\infty, -1] \cup [1, \infty)$.

An important point is that the class of infinitely divisible distributions is remarkably large. An indication is the fact that infinitely divisible distributions are characterized by the triples (b, σ^2, μ) , where μ is a measure on $\mathbb{R} - \{0\}$ satisfying (4.5). Two Lévy processes with triples (b_1, σ_1^2, μ_1) and (b_2, σ_2^2, μ_2) reduce to the same process if and only if $b_1 = b_2$, $\sigma_1^2 = \sigma_2^2$ and $\mu_1(A) = \mu_2(A)$ for all measurable sets $A \subseteq \mathbb{R}$. Nevertheless, infinitely divisible distributions may seem very special. However, over the years, many common distributions have been shown to be infinitely divisible. For example, lognormal distributions, Weibull distributions with ccdf's $e^{-(t/a)^c}$ for $c \leq 1$, Pareto distributions, and all mixtures of exponential distributions are infinitely divisible; see Thorin (1977a,b), p. 452 of Feller (1971), Bondesson (1992) and Abate and Whitt (1996). (The Weibull and Pareto distributions actually are mixtures of exponential distributions so infinite divisibility follows from that structure.) Moreover, the class of infinitely divisible distributions is easily seen to be closed under convolutions.

We now consider convergence in distribution of partial sums to infinitely divisible distributions and Lévy processes. First note that each infinitely divisible distribution can serve as a limit, because if X is infinitely divisible

then there is a sequence of sequences $\{X_{n,k} : k \geq 1\}$ of IID random variables such that $X \stackrel{d}{=} S_n$ for all n by the definition of infinite divisibility.

The following characterization of all possible limits is a consequence of Theorem 2, p. 303, of Feller (1971) and Theorem 2.7 of Skorohod (1957).

Theorem 2.4.1. (Lévy process FCLT for double sequences) *Let $\{X_{n,k} : k \geq 1\}$ be a sequence of IID random variables for each n and let $S_{n,n}$ and $\mathbf{S}_n$ be defined as in (4.1) and (4.2). If*

$$S_{n,n} \Rightarrow Z \quad \text{in } \mathbb{R} ,$$

then Z has an infinitely divisible distribution and

$$\mathbf{S}_n \Rightarrow \mathbf{L} \quad \text{in } D([0, \infty), J_1) ,$$

where $\mathbf{L}$ is the Lévy process with $\mathbf{L}(1) \stackrel{d}{=} Z$.

Necessary and sufficient conditions for the FCLT with convergence to a specific Lévy process are consequences of Theorems 2.35, 2.52 and 3.4 of pp. 362, 368 and 373 of Jacod and Shiryaev (1987). (The partial sum process is both a semimartingale and a process with independent increments (PII) but not a process with stationary independent increments (PIIS).)

Theorem 2.4.2. (criteria for the Lévy-process FCLT) *Let $\{X_{n,k} : k \geq 1\}$ be a sequence of IID random variables for each n , with $\{X_{n,1} : n \geq 1\}$ being infinitesimal, i.e.,*

$$\lim_{n \rightarrow \infty} P(|X_{n,1}| > \epsilon) = 0 \quad \text{for all } \epsilon > 0 . \quad (4.8)$$

Then

$$\mathbf{S}_n \Rightarrow \mathbf{L} \quad \text{in } D([0, \infty), \mathbb{R}, J_1) \quad (4.9)$$

for $\mathbf{S}_n$ in (4.2), where $\mathbf{L}$ is a Lévy process with characteristics (b, σ^2, μ) , if and only if

$$(i) \quad \lim_{n \rightarrow \infty} nE[h(X_{n,1})] = b , \quad (4.10)$$

$$(ii) \quad \lim_{n \rightarrow \infty} nVar[h(X_{n,1})] = \sigma^2 , \quad (4.11)$$

$$(iii) \quad \lim_{n \rightarrow \infty} nE[g(X_{n,1})] = \int_{-\infty}^{\infty} g(x)\mu(dx) , \quad (4.12)$$

for the truncation function h and all continuous bounded real-valued functions g on $\mathbb{R}$ with $g(x) = 0$ in a neighborhood of 0 and $g(x) \rightarrow y$, $-\infty < y < \infty$, as $x \rightarrow \pm\infty$.

Note that $h(x) = x$ for $|x| \leq 1$, so that conditions (i) and (ii) above correspond closely to convergence of the scaled means and variances.

Theorem 2.4.2 provides a large class of initial FCLT's to use with the continuous-mapping approach. We have only stated the classical results. Jacod and Shiryaev (1987) go much further, generalizing the characteristics of a Lévy process to define characteristics for semimartingales, allowing for nonstationarity. They also establish conditions for FCLTs in which processes with independent increments converge to other processes with independent increments (Chapter VII), semimartingales converge to processes with independent increments (Chapter VIII) and semimartingales converge to other semimartingales (Chapter IX), all expressed via the process characteristics. Actually verifying these conditions may not be straightforward, however.

2.5. Linear Models

In this section we discuss the linear-process representation in equation 4.6.6 of the book that was critical for obtaining the FCLT with strong dependence. The linear-process representation expresses the basic summands X_n as

$$X_n \equiv \sum_{j=0}^{\infty} a_j Y_{n-j}, \quad n \geq 1, \quad (5.1)$$

where $\{Y_n : -\infty < n < \infty\}$ is a two-sided sequence of IID random variables with $EY_n = 0$ and $EY_n^2 = 1$, and $\{a_j : j \geq 0\}$ is a sequence of (deterministic, finite) constants with

$$\sum_{j=0}^{\infty} a_j^2 < \infty. \quad (5.2)$$

We now show that the linear-process representation can arise naturally from modeling. First, however, it is important to repeat our earlier disclaimer. It is important to realize that the stochastic-process limits with strong dependence characterized by (5.1) are less universal. Many other forms of strong dependence are possible. And, if the dependence does not approximately correspond to a linear process, then there may appear a very different limit process or there may even be no stochastic-process limit at all.

Nevertheless, the linear-process representation is very natural. It provides a useful concrete model of strong dependence with an associated FCLT. To explain how linear models can arise, we describe some time-series models.

In particular, we show how the Gaussian linear process arises from a fundamental time-series model. We especially want to show how the Gaussian linear process with strong dependence arises from the *fractional autoregressive integrated moving average* (FARIMA) model; e.g., see Section 2.5 of Beran (1994) and Sections 7.12 and 7.13 of Samorodnitsky and Taqqu (1994).

The starting point is the *autoregressive moving average* (ARMA (p, q)) process, where p and q are nonnegative integers. To define the ARMA (p, q) process, let B be the *backshift operator*, defined by $BX_n \equiv X_{n-1}$, so that *differences* can be expressed as $X_n - X_{n-1} \equiv (1 - B)X_n$ and $(X_n - X_{n-1}) - (X_{n-1} - X_{n-2}) \equiv (1 - B)^2 X_n$. Let ϕ and ψ be polynomials of degree p and q , respectively, of the form

$$\phi(z) \equiv 1 - \sum_{j=1}^p \phi_j z^j$$

and

$$\psi(z) \equiv 1 + \sum_{j=1}^q \psi_j z^j,$$

where z is a complex variable and $\phi_1, \dots, \phi_p, \psi_1, \dots, \psi_q$ are real coefficients. As regularity conditions, assume that the equations $\phi(z) = 0$ and $\psi(z) = 0$ have no common roots and that all solutions of the equation $\phi(z) = 0$ fall outside the unit disk $\{z : |z| \leq 1\}$. An ARMA (p, q) process is defined to be the stationary solution to the equation

$$\phi(B)X_n = \psi(B)Y_n \tag{5.3}$$

where $\{Y_n : n \geq 1\}$ is a sequence of IID $N(0, 1)$ random variables; e.g., see Chapter 3 of Box, Jenkins and Reinsel (1994). In this setting, the sequence $\{Y_n\}$ is called the *innovation process*. Note that the exponential smoothing in Example 1.4.2 in the book is an ARMA(1, 0) process.

Theorem 2.5.1. (the ARMA process) *Under the regularity conditions above, the system of ARMA (p, q) equations (5.3) has a unique solution of the form*

$$X_n = \sum_{j=0}^{\infty} w_j Y_{n-j}, \quad n \geq 1, \tag{5.4}$$

with real constant coefficients w_j satisfying $|w_j| < \delta^j$ for all sufficiently large j , for some δ , $0 < \delta < 1$. The coefficients w_j in (5.4) are the coefficients of the power series $\psi(z)/\phi(z)$.

Note that the coefficients w_j in the linear-process representation are available via their generating function $\psi(z)/\phi(z)$. Given the polynomials ψ and ϕ , we can thus calculate the coefficients w_j by numerically inverting the generating function; see Abate and Whitt (1992b).

Also note that the coefficients w_j in (5.4) decay exponentially fast, so that an ARMA process only exhibits weak dependence. To obtain strong dependence, we need the coefficients w_j in (5.4) to decay more slowly. We achieve that by considering fractional differencing. We do so by introducing a generalization of the ARIMA model. If instead $\{X_n\}$ is the solution of the equation

$$\phi(B)(1 - B)^d X_n = \psi(B)Y_N, \quad (5.5)$$

where d is a nonnegative integer and $\{Y_n\}$ is again a sequence of IID $N(0, 1)$ random variables, then $\{X_n\}$ is said to be an ARIMA (p, d, q) process, which was introduced by Box and Jenkins (1970); see Chapter 4 of Box, Jenkins and Reinsel (1994).

The FARIMA process is a generalization of the ARIMA process to fractional differencing. The FARIMA generalization of ARIMA was introduced by Granger and Joyeux (1980) and Hosking (1981). The FARIMA model with strong dependence depends on a parameter triple (p, q, d) , where p and q are nonnegative integers and $0 < d < 1/2$. (There also are FARIMA models with $-1/2 < d \leq 0$, but we will not consider them.) Given (p, q) , there are $p + q$ further parameters.

For any real number d , we define the *fractional difference operator*

$$(1 - B)^d \equiv \sum_{k=0}^{\infty} \binom{d}{k} (-1)^k B^k,$$

where

$$\binom{d}{k} \equiv \frac{d!}{k!(d-k)!} \equiv \frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d-k+1)}$$

with $\Gamma(x)$ the gamma function. A stationary process $\{X_n\}$ that satisfies (5.5) for positive integers p and q and for $0 < d < 1/2$ is a FARIMA (p, d, q) process. (Values of d with $-1/2 < d \leq 0$ are also possible, but we are primarily interested in the range $0 < d < 1/2$.)

Theorem 2.5.2. (the FARIMA process) *Under the regularity conditions above, including $0 < d < 1/2$, the system of FARIMA (p, d, q) equations (5.5) has a unique solution of the form*

$$X_n = \sum_{j=0}^{\infty} a_j Y_{n-j}, \quad n \geq 1,$$

which converges almost surely, where

$$a_j \equiv \sum_{i=0}^j w_i b_{j-i}(-d)$$

with $\{w_j\}$ being the sequence of constant coefficients in (5.4) and

$$b_j(-d) \equiv \frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)} \sim \frac{1}{\Gamma(d)} j^{d-1} \quad \text{as } j \rightarrow \infty.$$

As a consequence,

$$a_j \sim a j^{d-1} \quad \text{as } j \rightarrow \infty,$$

where

$$a \equiv \sum_{i=0}^j w_i / \Gamma(d)$$

for w_i in (5.4), and

$$r_j \equiv \text{Cov}(X_1, X_{1+j}) \sim r j^{2d-1} \quad \text{as } j \rightarrow \infty,$$

where

$$r \equiv \left(\frac{\psi(1)}{\Gamma(d)\phi(1)} \right)^2 \int_0^\infty g(x) dx$$

for

$$g(x) = x^{2(d-1)} + (1+x)^{2(d-1)} - (x^{d-1} - (1+x)^{d-1})^2.$$

The point of this discussion has been to show that a linear process of the form (5.1) and (5.2), with

$$\text{Var}(S_n) = n^{2H} L(n) \quad \text{as } n \rightarrow \infty, \quad (5.6)$$

where $L(t)$ is a slowly varying function and $H > 1/2$, arises naturally from the FARIMA (p, d, q) model with $0 < d < 1/2$. In the FARIMA case the linear process is also a Gaussian process, but the key relations in Theorems 2.5.1 and 2.5.2 here hold for stationary sequences with finite second moments. We also remark that the parameters H and d are related by

$$d = H - \frac{1}{2}.$$

It is also significant that the FARIMA model provides a natural framework to exploit the strong dependence in order to make predictions; see

Beran (1994) for a full account of statistics for strongly dependent, light-tailed processes. We only make a few remarks.

In applications, we may have a stochastic sequence $\{X_n\}$ that we are willing to regard as a zero-mean stationary sequence with $\text{Var}(X_n) < \infty$. We can examine the variance $\text{Var}(S_n)$. If we find that

$$\text{Var}(S_n) \sim cn^{2H} \quad \text{as } n \rightarrow \infty$$

for $1/2 < H < 1$, then we have the Joseph effect. That can be checked by looking for a linear relationship after taking logarithms; i. e.,

$$\log(\text{Var}(S_n)) \sim \log(c) + 2H \log(n) .$$

We then can invoke Theorem 4.6.1 in the book, without directly verifying the linear-process representation in (5.1) and without identifying the weights a_j in (5.1), to support the approximation (in distribution)

$$\{(cn^{2H})^{-1/2} S_{[nt]} : t \geq 0\} \approx \{Z_H(t) : t \geq 0\} , \quad (5.7)$$

where Z_H is standard FBM. Note that we obtain a parsimonious approximation, depending only on the two parameters c and H . Attention naturally focuses on ways to estimate the parameters c and H . That can be done simply from a plot of $\log \text{Var}(S_n)$ as a function of $\log n$; see Beran (1994).

It is important to remember that the justification of approximation (5.7) from Theorem 4.6.1 in the book actually depends on the linear-process representation. However, we can directly justify the approximation equation 4.6.13 in the book, by checking that the finite-dimensional distributions are approximately Gaussian and that the covariance function is approximately the covariance function of FBM in equation 4.6.13 in the book. The limit theorem explains why the FBM approximation may be appropriate.

We conclude by remarking that there is again a time-series motivation for considering the linear-process representation in the case of heavy tails plus dependence, discussed in Section 4.7 of the book. Specifically, there is a time-series motivation for the linear-process representation in equation 4.7.1 of the book, where the innovation variables Y_n have heavy tails, just as there was for the light-tailed case in Section 4.6 of the book, because there are analogs of the ARMA, ARIMA and FARIMA processes with stable innovations; i.e., there are analogs of Theorems 2.5.1 and 2.5.2 here for the case in which the innovation process $\{Y_n\}$ is a sequence of IID random variables with a stable law $S_\alpha(\sigma, \beta, \mu)$ for $0 < \alpha < 2$; see Sections 7.12 and 7.13 of Samorodnitsky and Taqqu (1994).

Chapter 3

Preservation of Pointwise Convergence

3.1. Introduction

With the continuous-mapping approach to stochastic-process limits, we are concerned about limits $x_n \rightarrow x$ and $f_n(x_n) \rightarrow f(x)$ for a sequences of functions $\{x_n : n \geq 1\}$ in D and $f_n, f : D \rightarrow D$; see Section 3.5 and Chapter 13 in the book. However, in many applications we actually are interested in the pointwise limits

$$x(t)/\phi(t) \rightarrow \gamma \quad \text{in } \mathbb{R} \quad \text{as } t \rightarrow \infty \quad (1.1)$$

and

$$f(x)(t)/\phi(t) \rightarrow \eta \quad \text{in } \mathbb{R} \quad \text{as } t \rightarrow \infty \quad (1.2)$$

for single functions $x \in D$ and $f : D \rightarrow D$, where ϕ is a suitable scaling function. In particular, we may want to show that the pointwise limit (1.1) implies the associated pointwise limit (1.2) and identify the limit η .

It is significant that we often can obtain such limits in $\mathbb{R}$ as consequences of function-space limits by setting

$$y_s(t) \equiv x(st)/\phi(s), \quad s > 0. \quad (1.3)$$

As a regularity condition, we will assume that the scaling function ϕ is a homeomorphism of $\mathbb{R}_+$, i.e., $\phi \in \Lambda(\mathbb{R}_+)$. That implies that $\phi(0) = 0$, ϕ is increasing and $\phi(t) \rightarrow \infty$ as $t \rightarrow \infty$. If we can show that

$$y_s \rightarrow y \quad \text{in } D \quad \text{as } s \rightarrow \infty \quad (1.4)$$

for y_s in (1.3), where $1 \notin \text{Disc}(y)$, then we can apply the projection map π_1 taking x into $x(1)$ to obtain

$$y_s(1) = x(s)/\phi(s) \rightarrow y(1) \quad \text{in } \mathbb{R} \quad \text{as } s \rightarrow \infty, \quad (1.5)$$

which implies the desired convergence in (1.1) and identifies the limit γ in (1.1) as $y(1)$ in (1.5). Moreover, if we can show that

$$f(x)(t)/\phi(t) = g_t(y_t) \quad \text{for each } t > 0, \quad (1.6)$$

where $g_s, g : D \rightarrow \mathbb{R}$ and

$$g_s(y_s) \rightarrow g(y) \quad \text{in } \mathbb{R} \quad (1.7)$$

whenever $y_s \rightarrow y$ in D , then we can obtain (1.2) from (1.4) as well, and we identify the limit η in (1.2) as $g(y)$.

For example, this reasoning applies to the supremum function: $f(x)(t) = x^\uparrow(t)$ for $t > 0$. Then $g(y) = g_s(y) = f(y)(1) = y^\uparrow(1)$ for all $y \in D$ and $s > 0$. As a consequence, the limit in (1.2) holds with $\eta = g(y) = y^\uparrow(1)$.

Even though many pointwise limits for single functions can be subsumed as special cases of function-space limits, it is interesting to consider what can be obtained directly without resorting to the function-space construction in (1.3). In particular, it is natural to ask how pointwise limits for single functions are preserved under the composition, supremum, and inverse maps. We investigate that question in this chapter.

For queues and related applied probability models, this convergence-preservation issue for single sample paths corresponds to sample-path analysis, which is commonly associated with the fundamental relations (conservation laws) $L = \lambda W$ and Arrivals See Time Averages (ASTA); see El-Taha and Stidham (1999); see the chapter notes at the end of the chapter.

3.2. From Pointwise to Uniform Convergence

Clearly, the pointwise limit in (1.1) is more elementary than the function-space limit (1.4) but, surprisingly, (1.4) is not much stronger than (1.1). Indeed, under minor regularity conditions, (1.1) actually implies (1.4). Recall that ϕ in $\Lambda(\mathbb{R}_+)$ is regularly varying with index $p > 0$, denoted by $\phi \in \mathcal{R}(p)$, if

$$\phi(tx)/\phi(t) \rightarrow x^p \quad \text{as } t \rightarrow \infty \quad (2.1)$$

for all $x > 0$; see Appendix A at the end of the book.

Theorem 3.2.1. (from pointwise to uniform convergence) *Let $x \in D$ and $\phi \in \Lambda(\mathbb{R}_+)$ with $\phi \in \mathcal{R}(p)$ for $p > 0$. If the limit (1.1) holds in $\mathbb{R}$, then*

$$\|y_s - y\|_T \rightarrow 0 \quad \text{as } s \rightarrow \infty \quad \text{for each } T > 0 \quad (2.2)$$

for y_s in (1.3) and

$$y(t) = \gamma t^p, \quad t \geq 0.$$

Proof. Under the conditions, for any $\epsilon > 0$, there is a t_0 such that

$$|x(t)/\phi(t) - \gamma| < \epsilon \quad \text{for all } t \geq t_0. \quad (2.3)$$

and an s_0 such that

$$\sup_{0 \leq t \leq T} \left| \frac{\phi(st)}{\phi(s)} - t^p \right| < \epsilon \quad \text{for all } s \geq s_0; \quad (2.4)$$

see Theorem A.5 in Appendix A in the book. For $t \leq t_0/s$,

$$|y_s(t) - y(t)| \leq |y_s(t)| + |y(t)| \leq \frac{\|x\|_{t_0}}{\phi(s)} + \gamma \left(\frac{t_0}{s} \right)^p, \quad (2.5)$$

which is less than ϵ for all sufficiently large s , say $s \geq s_1 \geq s_0$. Since

$$y_s(t) - y(t) = \frac{x(st)}{\phi(st)} \left(\frac{\phi(st)}{\phi(s)} - t^p \right) + t^p \left(\frac{x(st)}{\phi(st)} - \gamma \right), \quad (2.6)$$

for $s \geq s_1$,

$$\begin{aligned} \|y_s - y\|_T &\leq \epsilon + \sup_{t \geq t/s} \left\{ \left| \frac{x(st)}{\phi(st)} \right| \left| \frac{\phi(st)}{\phi(s)} - t^p \right| + t^p \left| \frac{x(st)}{\phi(st)} - \gamma \right| \right\} \\ &\leq \epsilon + (\gamma + \epsilon)\epsilon + T^p \epsilon, \end{aligned} \quad (2.7)$$

which can be made arbitrarily small with an appropriate choice of ϵ . ■

For the special case in which $\phi(t) = t$, condition (1.1) corresponds to a strong law of large numbers (SLLN) for a stochastic process, while the conclusion (2.2) corresponds to a functional strong law of large numbers (FSLLN). The following corollary is Theorem 4 from Glynn and Whitt (1988).

Corollary 3.2.1. (from a SLLN to a FSLLN) *Let $\{X(t) : t \geq 0\}$ be a real-valued stochastic process and let*

$$\hat{X}_n(t) \equiv n^{-1}X(nt), \quad t \geq 0, \quad n \geq 1. \quad (2.8)$$

If a SLLN holds, i.e., if

$$t^{-1}X(t) \rightarrow \gamma \text{ w.p.1 in } \mathbb{R} \text{ as } t \rightarrow \infty, \quad (2.9)$$

then a FSLLN holds, i.e.,

$$\|\hat{\mathbf{X}}_n - \gamma \mathbf{e}\|_T \rightarrow 0 \text{ w.p.1 in } D([0, T], \mathbb{R}) \text{ as } n \rightarrow \infty \quad (2.10)$$

for all $T > 0$.

3.3. Supremum

In this section we consider the supremum map. The following elementary convergence-preservation result is referred to as the “fundamental lemma of maxima” in Section 2.5 of El-Taha and Stidham (1999).

Proposition 3.3.1. (preservation of pointwise convergence for the supremum) *Suppose that $x \in D([0, \infty), \mathbb{R})$, ϕ is an increasing real-valued function on $\mathbb{R}_+$ with $\phi(t) \rightarrow \infty$ as $t \rightarrow \infty$. If $x(t)/\phi(t) \rightarrow \gamma \geq 0$ as $t \rightarrow \infty$, then $x^\uparrow(t)/\phi(t) \rightarrow \gamma$ as $t \rightarrow \infty$.*

Proof. Under the condition, for any $\epsilon > 0$, there exists t_0 such that

$$(\gamma - \epsilon)\phi(t) \leq x(t) \leq (\gamma + \epsilon)\phi(t)$$

for all $t \geq t_0$. Hence,

$$(\gamma - \epsilon)\phi(t) \leq x^\uparrow(t) \leq x^\uparrow(t_0) \vee (\gamma + \epsilon)\phi(t)$$

for all $t \geq t_0$. Since $\gamma \geq 0$ and $\phi(t) \rightarrow \infty$, there is $t_1 \geq t_0$ such that $x^\uparrow(t_0) \leq (\gamma + \epsilon)\phi(t)$ for all $t \geq t_1$. Thus, for $t \geq t_1$,

$$|\phi(t)^{-1}x^\uparrow(t) - \gamma| \leq \epsilon. \quad \blacksquare$$

Under the conditions of Theorem 3.2.1, if $\gamma \geq 0$, then we can apply the continuous mapping theorem to deduce that $x^\uparrow(t)/\phi(t) \rightarrow \gamma$ as $t \rightarrow \infty$; i.e., the conclusion of Proposition (3.3.1) holds by virtue of Theorems 3.2.1 here and 13.4.1 in the book. However, Theorem 3.2.1 here has the extra assumption that ϕ is regularly varying.

Paralleling Proposition (3.3.1), we can also establish a pointwise-convergence result for supremum with centering for a single function.

Proposition 3.3.2. (preservation of pointwise convergence with centering for the supremum) *Suppose that ϕ is an increasing real-valued function such that $\phi(t) \rightarrow \infty$ and $\phi(t)/t \rightarrow 0$ as $t \rightarrow \infty$. If*

$$\phi(t)^{-1}[x(t) - \lambda t] \rightarrow \gamma \quad \text{as } t \rightarrow \infty \quad (3.1)$$

for $\lambda > 0$, then

$$\phi(t)^{-1}[x^\uparrow(t) - \lambda t] \rightarrow \gamma \quad \text{as } t \rightarrow \infty. \quad (3.2)$$

Proof. Under condition (3.1), for any $\epsilon > 0$, there exists t_0 such that

$$\lambda t - \phi(t)(\gamma - \epsilon) \leq x(t) \leq \lambda t + \phi(t)(\gamma + \epsilon)$$

for all $t \geq t_0$. Then

$$\lambda t - \phi(t)(\gamma - \epsilon) \leq x^\uparrow(t) \leq x^\uparrow(t_0) \vee (\lambda t + \phi(t)(\gamma + \epsilon)).$$

However, since $\lambda > 0$ and $\phi(t)/t \rightarrow 0$, there is a $t_1 > t_0$ such that $x^\uparrow(t_0) \leq \lambda t + \phi(t)(\gamma + \epsilon)$ for all $t \geq t_1$. Hence

$$|\phi(t)^{-1}[x^\uparrow(t) - \lambda t] - \gamma| < \epsilon$$

for all $t \geq t_1$, so that (3.2) holds. ■

3.4. Counting Functions

We now turn to counting functions, as in Section 13.8 of the book. A counting function is defined in terms of a sequence $\{s_n : n \geq 0\}$ of nondecreasing nonnegative real numbers with $s_0 = 0$. We can think of s_n as the partial sum

$$s_n \equiv x_1 + \cdots + x_n, \quad n \geq 1, \quad (4.1)$$

by simply writing $x_i \equiv s_i - s_{i-1}$, $i \geq 1$. The associated *counting function* $\{c(t) : t \geq 0\}$ is defined by

$$c(t) \equiv \max\{k \geq 0 : s_k \leq t\}, \quad t \geq 0. \quad (4.2)$$

To have $c(t)$ finite for all $t > 0$, we assume that $s_n \rightarrow \infty$ as $n \rightarrow \infty$.

To establish limits for counting functions, we use two scaling functions. We again let the scaling functions be elements of $\Lambda(\mathbb{R}_+)$. Note that if $\phi \in \Lambda(\mathbb{R}_+)$, then $\phi(0) = 0$ and ϕ is strictly increasing. Also, ϕ necessarily has an inverse ϕ^{-1} with $\phi \circ \phi^{-1} = \phi^{-1} \circ \phi = e$. Moreover, $(\phi_1 \circ \phi_2)^{-1} = \phi_2^{-1} \circ \phi_1^{-1}$ for two homeomorphisms ϕ_1 and ϕ_2 .

The basis for positive results is the basic inverse relation in Lemma 13.8.1 of the book, which we restate here:

Lemma 3.4.1. (basic inverse relation) *For any nonnegative integer n and nonnegative real number t ,*

$$s_n \leq t \quad \text{if and only if} \quad c(t) \geq n. \quad (4.3)$$

The relation between the limits for s_n as $n \rightarrow \infty$ and $c(t)$ as $t \rightarrow \infty$ follows easily from the following bounds, which are of independent interest. Let $\lfloor x \rfloor$ be the greatest integer less than or equal to x and let $\lceil x \rceil$ be the least integer greater than or equal to x . One-sided bounds are obtained below by either setting $\epsilon = 1$ or setting $\delta = \infty$. Let $1/0 = \infty$ and $1/\infty = 0$.

Lemma 3.4.2. (one-sided bounds) *Suppose that $\phi_1, \phi_2 \in \Lambda(\mathbb{R}_+)$, $0 < \epsilon \leq 1$ and $0 < \delta \leq \infty$.*

(a) If

$$1 - \epsilon \leq \frac{\phi_2(c(t))}{\phi_1(t)} < 1 + \delta \quad \text{for all} \quad t \geq t_0, \quad (4.4)$$

then

$$\frac{1}{1 + \delta} < \frac{\phi_1(s_n)}{\phi_2(n)} \leq \frac{1}{1 - \epsilon} \quad \text{for all} \quad n \geq n_0 \equiv \lceil \phi_2^{-1}(\phi_1(t_0)(1 + \delta)) \rceil. \quad (4.5)$$

(b) If

$$1 - \epsilon < \frac{\phi_1(s_n)}{\phi_2(n)} \leq 1 + \delta \quad \text{for all} \quad n \geq n_0, \quad (4.6)$$

then

$$\frac{\phi_2(c(t))}{\phi_1(t)} \leq \frac{1}{1 - \epsilon} \quad (4.7)$$

and

$$\frac{\phi_2(c(t) + 1)}{\phi_1(t)} \geq \frac{1}{1 + \delta}. \quad (4.8)$$

for all $t \geq t_0 \equiv \lceil \phi_1^{-1}(\phi_2(t_0)(1 + \delta)) \rceil$. Moreover, there is a sequence of times $\{t_k\}$ such that $t_k \rightarrow \infty$ as $k \rightarrow \infty$ and

$$\frac{\phi_2(c(t_k))}{\phi_1(t_k)} \geq \frac{1}{1 + \delta} \quad (4.9)$$

for all $t_k \geq t_0$.

Proof. (a) If (4.4) holds, then

$$n_1(t) \equiv \lfloor \phi_2^{-1}(\phi_1(t)(1 - \epsilon)) \rfloor \leq c(t) < \lceil \phi_2^{-1}(\phi_1(t)(1 + \delta)) \rceil \equiv n_2(t)$$

for all $t \geq t_0$ and, by Lemma 3.4.1,

$$s_{n_1(t)} \leq t < s_{n_2(t)} \quad \text{for all } t \geq t_0. \quad (4.10)$$

Let t_1 and t_2 be functions of n defined by

$$t_1(n) \equiv \phi_1^{-1}(\phi_2(n)/(1 - \epsilon)) \quad \text{and} \quad t_2(n) \equiv \phi_1^{-1}(\phi_2(n)/(1 + \delta)),$$

and note that $n_1(t_1(n)) = n_2(t_2(n)) = n$ for all n . Hence, for all $n \geq n_0 \equiv \lceil \phi_2^{-1}(\phi_1(t_0)(1 + \delta)) \rceil$, we have $t_1(n_0) \geq t_2(n_0) \geq t_0$ and, by (4.10),

$$t_2(n) < s_{n_2(t_2(n))} = s_n = s_{n_1(t_1(n))} \leq t_1(n)$$

or, equivalently,

$$\phi_2(n) \left(\frac{1}{1 + \delta} - 1 \right) < \phi_1(s_n) - \phi_2(n) \leq \phi_2(n) \left(\frac{1}{1 - \epsilon} - 1 \right)$$

which implies (4.5).

(b) If (4.6) holds, then

$$\tilde{t}_1(n) \equiv \phi_1^{-1}(\phi_2(n)(1 - \epsilon)) < s_n \leq \phi_1^{-1}(\phi_2(n)(1 + \delta)) \equiv \tilde{t}_2(n)$$

for all $n \geq n_0$ and, by Lemma 3.4.1,

$$c(\tilde{t}_1(n)) < n \leq c(\tilde{t}_2(n)) \quad \text{for all } n \geq n_0. \quad (4.11)$$

Let $\tilde{n}_1$ and $\tilde{n}_2$ be functions of t defined by

$$\tilde{n}_1(t) \equiv \lceil \phi_2^{-1}(\phi_1(t)/(1 - \epsilon)) \rceil \quad \text{and} \quad \tilde{n}_2(t) \equiv \lfloor \phi_2^{-1}(\phi_1(t)/(1 + \delta)) \rfloor$$

and note that

$$\tilde{t}_2(\tilde{n}_2(t)) \leq t \leq \tilde{t}_1(\tilde{n}_1(t)).$$

Hence, by (4.11),

$$\tilde{n}_2(t) \leq c(\tilde{t}_2(\tilde{n}_2(t))) \leq c(t) \leq c(\tilde{t}_1(\tilde{n}_1(t))) < \tilde{n}_1(t)$$

and

$$\phi_2^{-1}(\phi_1(t)/(1 + \delta)) - 1 \leq c(t) \leq \phi_2^{-1}(\phi_1(t)/(1 - \epsilon))$$

for all $t \geq t_0 \equiv \phi_1^{-1}(\phi_2(n_0)(1 + \delta))$, because $\tilde{n}_1(t_0) \geq \tilde{n}_2(t_0) = n_0$, which implies (4.7) and (4.8) by the reasoning for part (a). For (4.9), choose the sequence $\{t_k\}$ so that $\phi_2^{-1}(\phi_1(t_k)/(1 + \delta))$ is an integer. Then we have the lower bound $c(t_k) \geq \phi_2^{-1}(\phi_1(t_k)/(1 + \delta))$ for all k , which implies (4.9). ■

We now apply Lemma 3.4.2 to characterize the asymptotic behavior.

Theorem 3.4.1. (implications for pointwise convergence) *Suppose that $\phi_1, \phi_2 \in \Lambda(\mathbb{R}_+)$ and $0 \leq \lambda \leq \infty$.*

(a) If $\phi_2(c(t))/\phi_1(t) \rightarrow \lambda$ as $t \rightarrow \infty$, then $\phi_1(s_n)/\phi_2(n) \rightarrow \lambda^{-1}$ as $n \rightarrow \infty$.

(b) If $\phi_1(s_n)/\phi_2(n) \rightarrow \lambda^{-1}$ as $n \rightarrow \infty$, then

$$\overline{\lim}_{t \rightarrow \infty} \phi_2(c(t))/\phi_1(t) = \lambda. \quad (4.12)$$

(c) If, in addition to the condition for (b), either

$$\frac{\phi_2(c(t) + 1) - \phi_2(c(t))}{\phi_1(t)} \rightarrow 0 \quad \text{as } t \rightarrow \infty \quad (4.13)$$

or

$$\frac{\phi_2(n + 1)}{\phi_2(n)} \rightarrow 1 \quad \text{as } n \rightarrow \infty, \quad (4.14)$$

then $\phi_2(c(t))/\phi_1(t) \rightarrow \lambda$ as $t \rightarrow \infty$.

(d) If $\phi_1(s_n)/\phi_2(n) \rightarrow 0$ as $n \rightarrow \infty$ and either

$$\overline{\lim}_{t \rightarrow \infty} \frac{\phi_2(c(t) + 1) - \phi_2(c(t))}{\phi_1(t)} < \infty \quad (4.15)$$

or

$$\underline{\lim}_{n \rightarrow \infty} \frac{\phi_2(n)}{\phi_2(n + 1)} > 0, \quad (4.16)$$

then $\phi_2(c(t))/\phi_1(t) \rightarrow \infty$ as $t \rightarrow \infty$.

Proof. (a) First suppose that $0 < \lambda < \infty$. Then incorporate λ into $\phi_1(t)$ by dividing by λ . The condition implies that for all appropriate ϵ and δ there exists t_0 such that (4.4) holds. By Lemma 3.4.2(a), (4.5) holds. Since ϵ and δ are arbitrary in (4.5), it implies the desired conclusion. To treat the cases $\lambda = 0$ and $\lambda = \infty$, use the one-sided bounds in Lemma 3.4.2. For example, if $\phi_2(c(t))/\phi_1(t) \rightarrow 0$ as $t \rightarrow \infty$, then for all positive ϵ and δ there exists t_0 such that $\phi_2(c(t))/\epsilon\phi_1(t) < 1 + \delta$ for all $t \geq t_0$. By Lemma 3.4.2(a), $\epsilon\phi_1(s_n)/\phi_2(n) > 1/(1 + \delta)$ for all $n \geq n_0$. Since ϵ can be arbitrarily small, $\phi_1(s_n)/\phi_2(n) \rightarrow \infty$ as $n \rightarrow \infty$.

(b) Reason as in (a) using (4.6), (4.7) and (4.9).

(c) Use (4.8), (4.13) and (4.14), noting that

$$\frac{1}{1 - \epsilon} - \frac{\phi_2(c(t) + 1) - \phi_2(c(t))}{\phi_1(t)} \leq \frac{\phi_2(c(t))}{\phi_1(t)} \leq \frac{1}{1 - \epsilon} \quad (4.17)$$

and

$$\frac{\phi_2(c(t))}{\phi_2(c(t) + 1)(1 + \epsilon)} \leq \frac{\phi_2(c(t))}{\phi_1(t)} \leq \frac{1}{1 - \epsilon}. \quad (4.18)$$

(d) Reason as in (c), using (4.15) and (4.16) with (4.17) and (4.18). ■

Remark 3.4.1. Note that $\phi_2(c(t))/\phi_1(t) \rightarrow \lambda$ as $t \rightarrow \infty$ if and only if $\phi_2(c(\phi_1^{-1}(t)))/t \rightarrow \lambda$ as $t \rightarrow \infty$; i.e., the spatial normalization $\phi_1(t)$ is equivalent to the standard normalizing function e after making a time transformation by ϕ^{-1} . ■

Example 3.4.1. *The need for an extra condition.* To see that an extra condition is needed in Theorem 3.4.1(c), let $s_n = n$ for all n , so that $c(t) = \lfloor t \rfloor$ for all t . Also let $\phi_1(t) = \phi_2(t) = e^t$ for all t . Then $\phi_1(s_n)/\phi_2(n) = 1$ for all n , while

$$\phi_2(c(t))/\phi_1(t) = e^{\lfloor t \rfloor - t},$$

which has limit supremum 1 and limit infimum e^{-1} . Also note that neither (4.13) nor (4.14) is satisfied.

Example 3.4.2. *The extra conditions are not necessary.* To see that the specific extra conditions in Theorem 3.4.1(c) are not necessary, let $s_n = e^n$ for all n , so that $c(t) = \lfloor \log t \rfloor$. Let $\phi_2(t) = e^t$ and $\phi_1(t) = t$ for all t . Then $\phi_1(s_n)/\phi_2(n) = 1$ for all n and

$$\frac{\phi_2(c(t))}{\phi_1(t)} = \frac{e^{\lfloor \log t \rfloor}}{t} \rightarrow 1 \quad \text{as } t \rightarrow \infty,$$

but $\phi_2(n+1)/\phi_2(n) = e$ for all n and

$$\frac{\phi_2(c(t) + 1) - \phi_2(c(t))}{\phi_1(t)} = \frac{(e - 1)e^{\lfloor \log t \rfloor}}{t} \rightarrow e - 1 \quad \text{as } t \rightarrow \infty. \quad \blacksquare$$

A special case of interest is when the homeomorphisms are of the form $\phi(t) = t^p$ for $p > 0$. Of course, the case of greatest interest is $p = 1$; then we have simple averages.

Corollary 3.4.1. (the special case of powers) *Suppose that $0 < p < \infty$ and $0 \leq \lambda \leq \infty$. The following are equivalent:*

- (i) $c(t)/t^p \rightarrow \lambda$ as $t \rightarrow \infty$,
- (ii) $(c(t))^{1/p}/t \rightarrow \lambda^{1/p}$ as $t \rightarrow \infty$,
- (iii) $s_n/n^{1/p} \rightarrow \lambda^{-1/p}$ as $n \rightarrow \infty$,
- (iv) $(s_n)^p/n \rightarrow \lambda^{-1}$ as $n \rightarrow \infty$.

Proof. Apply Theorem 3.4.1 with $\phi_2(t) = t$ and $\phi_1(t) = t^p$ to relate (i) and (iv). Note that (4.14) holds. To relate (i) and (ii), note that $(c(t))^{1/p}/t = (c(t)/t^p)^{1/p}$, and similarly for (iii) and (iv). ■

We used the property that $\phi(x/y) = \phi(x)/\phi(y)$ for $\phi(x) = x^p$ in Corollary 3.4.1. The following classic lemma shows that this does not hold more generally.

Lemma 3.4.3. *A homeomorphism ϕ of $\mathbb{R}_+$ satisfies $\phi(xy) = \phi(x)\phi(y)$ for all nonnegative x and y if and only if $\phi(t) = t^p$ for some $p > 0$.*

Proof. The sufficiency is immediate. For the necessity, suppose that $\phi(xy) = \phi(x)\phi(y)$ for all nonnegative x and y . If we let $\psi(x) = \log \phi(e^x)$, then $\psi(x+y) = \psi(x) + \psi(y)$ for all real x and y . It is well known and easy to see that $\psi(x) = px$ for some real number p , which implies that $\phi(x) = e^{\psi(\log x)} = e^{p \log x} = x^p$. Since ϕ is strictly increasing, we must have $p > 0$. ■

The Corollary to Theorem 3.4.1 is useful because it enables us to replace $\phi_2(c(t))/\phi_1(t)$ and $\phi_1(s_n)/\phi_2(n)$ by $c(t)/\phi_2^{-1}(\phi_1(t))$ and $s_n/\phi_1^{-1}(\phi_2(n))$ respectively. The following lemma shows that we can do this more generally.

Lemma 3.4.4. *Suppose that $\phi \in \Lambda(\mathbb{R}_+)$, $a_n \rightarrow \infty$ and $a_n/b_n \rightarrow 1$ as $n \rightarrow \infty$. If there is a t_0 such that $\log \phi(e^t)$ is uniformly continuous in (t_0, ∞) , then $\phi(a_n)/\phi(b_n) \rightarrow 1$ as $n \rightarrow \infty$.*

Proof. Since $a_n \rightarrow \infty$ and $a_n/b_n \rightarrow 1$ as $n \rightarrow \infty$, $\log a_n - \log b_n \rightarrow 0$, $\log a_n \rightarrow \infty$ and $\log b_n \rightarrow \infty$ as $n \rightarrow \infty$. If $\log(\phi(e^t))$ is uniformly continuous in (t_0, ∞) , then

$$\begin{aligned} \log \phi(e^{\log a_n}) - \log \phi(e^{\log b_n}) &= \log \phi(a_n) - \log \phi(b_n) \\ &= \log(\phi(a_n)/\phi(b_n)) \rightarrow 0 \quad \text{as } n \rightarrow \infty, \end{aligned}$$

so that $\phi(a_n)/\phi(b_n) \rightarrow 1$ as $n \rightarrow \infty$. ■

The following Corollary to Lemma 3.4.4 indicates how Lemma 3.4.4 can be applied in our context.

Corollary 3.4.2. *If $\phi_2(c(t))/\phi_1(t) \rightarrow \lambda$ as $t \rightarrow \infty$, where $\phi_1, \phi_2 \in \Lambda(\mathbb{R}_+)$ and $\log \phi_2^{-1}(e^t)$ is uniformly continuous in (t_0, ∞) for some t_0 , then $c(t)/\phi_2^{-1}(\lambda\phi_1(t)) \rightarrow 1$ as $t \rightarrow \infty$.*

Remark 3.4.2. Lemma 3.4.4 implies Corollary 3.4.1 because $\log \phi(e^t) = \log \lambda + pt$ when $\phi(t) = \lambda t^p$. Another function covered by Lemma 3.4.4 is

$\phi(t) = a \log bt$; then $\log \phi(e^t) = \log a + \log(\log b + t)$. However, $\log \phi(e^t) = \log a + b e^t$ when $\phi(t) = a e^{bt}$, so that the uniform continuity does not hold when $\phi(t) = a e^{bt}$. ■

The following result is also useful to characterize the normalizing functions.

Lemma 3.4.5. *Suppose that $\phi \in \Lambda(\mathbb{R}_+)$, $0 < \lambda < \infty$ and $a_n \rightarrow \infty$ as $n \rightarrow \infty$. If there is a t_0 such that $\log \phi(e^t)$ is uniformly continuous in (t_0, ∞) , then*

$$\overline{\lim}_{n \rightarrow \infty} \left| \frac{\phi(a_n)}{\phi(\lambda a_n)} \right| < \infty.$$

Proof. Recall that if a function ψ is uniformly continuous in (t_0, ∞) , then

$$\sup\{|\psi(t+x) - \psi(t)| : t \geq t_0\} < \infty$$

for any positive x . Since

$$\log \lambda a_n - \log a_n = \log \lambda ,$$

$$\begin{aligned} \overline{\lim}_{n \rightarrow \infty} \{|\log \phi(e^{\log \lambda a_n}) - \log \phi(e^{a_n})|\} &= \overline{\lim}_{n \rightarrow \infty} \{|\log \phi(\lambda a_n) - \log \phi(a_n)|\} \\ &= \overline{\lim}_{n \rightarrow \infty} \{|\log(\phi(\lambda a_n)/\phi(a_n))|\} < \infty , \end{aligned}$$

which implies the desired conclusion. ■

We are thinking of $\{s_n : n \geq 1\}$ being the points in a point process sample path, so it is natural to assume that $\{s_n\}$ is nondecreasing. However, we could start with a general sequence of real numbers $\{t_n : n \geq 1\}$ and obtain $\{s_n\}$ as the successive maxima, i.e.,

$$s_n \equiv t_n^\uparrow \equiv \max\{t_k : 0 \leq k \leq n\}, \quad n \geq 1, \quad (4.19)$$

where $t_0 = 0$. A similar result holds for $c(t)$. The following result closely parallel Proposition 3.3.1.

Proposition 3.4.1. *Suppose that $\phi_1, \phi_2 \in \Lambda(\mathbb{R}_+)$ and $0 \leq \lambda \leq \infty$. If $\phi_1(t_n)/\phi_2(n) \rightarrow \lambda^{-1}$ as $n \rightarrow \infty$, then $\phi_1(s_n)/\phi_2(n) \rightarrow \lambda^{-1}$ as $n \rightarrow \infty$ for s_n in (4.19).*

Proof. First assume that $0 < \lambda < \infty$. Given the assumed convergence, for all $\epsilon > 0$, there is an n_0 such that

$$\phi_1^{-1}(\phi_2(n)/\lambda(1+\epsilon)) \leq t_n \leq \phi_1^{-1}(\phi_2(n)/\lambda(1-\epsilon)) \quad \text{for all } n \geq n_0,$$

which implies

$$\phi_1^{-1}(\phi_2(n)/\lambda(1+\epsilon)) \leq s_n \leq \max\{s_{n_0}, \phi_1^{-1}(\phi_2(n)/\lambda(1-\epsilon))\} \quad \text{for all } n \geq n_0.$$

Let n_1 be such that

$$\phi_1^{-1}(\phi_2(n)/\lambda(1-\epsilon)) \geq s_{n_0}.$$

Then, for all $n \geq n_1$,

$$\frac{1}{\lambda(1+\epsilon)} \leq \frac{\phi_1(s_n)}{\phi_2(n)} \leq \frac{1}{\lambda(1-\epsilon)},$$

which implies the conclusion. For $\lambda = 0$ and $\lambda = \infty$ use associated one-sided inequalities. ■

3.5. Counting Functions with Centering

We now turn to counting functions with centering. Due to the results for the inverse map with centering in Section 13.7 of the book, Theorem 13.8.2 in the book yields FCLTs for stochastic counting processes with centering given FCLTs for associated sequences of nondecreasing nonnegative random variables with centering, by an application of the continuous mapping theorem. We now show that we can also exploit the monotonicity to obtain *ordinary* CLTs for stochastic counting processes from associated *ordinary* CLTs for sequences of nondecreasing nonnegative random variables. The resulting CLT for stochastic counting process is the same as can be obtained from the FCLT by projection, but the condition is weaker. In both cases, we rely on an existing limit rather than specific stochastic assumptions. For this purpose, let $\{S_n : n \geq 0\}$ be a sequence of nondecreasing nonnegative random variables with $S_0 = 0$ and let $\{C(t) : t \geq 0\}$ be the associated stochastic counting process, defined as before by

$$C(t) \equiv \max\{k \geq 0 : S_k \leq t\}, \quad t \geq 0. \quad (5.1)$$

We again use regularly varying functions.

Theorem 3.5.1. (CLT equivalence) *Suppose that $m > 0$ and $\psi \in \mathcal{R}(p)$ for $0 < p < 1$. Then*

$$\psi(n)^{-1}[S_n - nm] \Rightarrow L \quad \text{in } \mathbb{R} \quad \text{as } n \rightarrow \infty, \quad (5.2)$$

where $\{S_n : n \geq 0\}$ is a sequence of nondecreasing nonnegative random variables with $S_0 = 0$ if and only if

$$\psi(t)^{-1}[C(t) - m^{-1}t] \Rightarrow -m^{-(1+p)}L \quad \text{in } \mathbb{R} \quad \text{as } n \rightarrow \infty, \quad (5.3)$$

where $\{C(t) : t \geq 0\}$ is the associated stochastic counting process.

We obtain Theorem 3.5.1 from a more general theorem which allows more general scalings, which are of value when analyzing nonstationary point processes.

Theorem 3.5.2. (CLT implications with more general scaling functions) *Suppose that $\phi_1, \phi_2 \in \Lambda(\mathbb{R}_+)$, $\psi \in C_\uparrow$ and*

$$\psi(t)/\psi(t + x\psi(t)) \rightarrow 1 \quad \text{as } t \rightarrow \infty \quad (5.4)$$

for all x .

(a) *If*

$$X(t) \equiv \psi(\phi_1(t))^{-1}[\phi_2(C(t)) - \phi_1(t)] \Rightarrow L \quad \text{in } \mathbb{R} \quad \text{as } n \rightarrow \infty, \quad (5.5)$$

then

$$Y(n) = \psi(\phi_2(n))^{-1}[\phi_1(S_n) - \phi_2(n)] \Rightarrow -L \quad \text{in } \mathbb{R} \quad \text{as } n \rightarrow \infty. \quad (5.6)$$

(b) *If (5.6) above holds, then there exists an increasing sequence of positive real numbers $\{t_n : n \geq 1\}$ with $t_n \rightarrow \infty$ as $n \rightarrow \infty$ such that $X(t_n) \Rightarrow L$ for $X(t)$ in (5.5) above.*

(c) *If, in addition to (5.6) above,*

$$[\phi_2(n+1) - \phi_2(n)]/\psi(\phi_2(n)) \rightarrow 1 \quad \text{as } n \rightarrow \infty \quad (5.7)$$

and

$$\psi(\phi_2(n+1))/\psi(\phi_2(n)) \rightarrow 1 \quad \text{as } n \rightarrow \infty, \quad (5.8)$$

then (5.5) above holds.

We first apply Theorem 3.5.2 to prove Theorem 3.5.1.

Proof of Theorem 3.5.1. We apply Theorem 3.5.2 with $\phi_1(t) = mt$, $\phi_2(t) = t$ and $\psi \in \mathcal{R}(p)$ for $0 < p < 1$. It is easy to see that ψ satisfies (5.4): For any x , there is a t_0 such that

$$\psi((1 - \epsilon)t) \leq \psi(t + x\psi(t)) \leq \psi((1 + \epsilon)t) \quad (5.9)$$

for all $t \geq t_0$, from which it follows that

$$\left(\frac{1}{1 + \epsilon}\right)^p \leq \frac{\psi(t)}{\psi(t + x\psi(t))} \leq \left(\frac{1}{1 - \epsilon}\right)^p \quad (5.10)$$

for all suitably large t . We also apply the regular variation property to deduce that $\psi(\phi_1(t))$ in (5.5) has the asymptotic form

$$\psi(\phi_1(t)) = \psi(mt) \sim m^p \psi(t) \quad \text{as } t \rightarrow \infty. \quad (5.11)$$

Thus (5.2) is equivalent to (5.6) with the limit in (5.2) changed to $m^p L$. Similarly, (5.3) is equivalent to (5.5) with the limit in (5.3) changed to $-m^{-1}L$. Thus the form of the limits in (5.2) and (5.3) follow from (5.5) and (5.6). Finally, it remains to observe that the assumptions in Theorem 3.5.1 imply that conditions (5.7) and (5.8) hold. ■

We now turn to the proof of Theorem 3.5.2. For that purpose, we use a basic lemma about cumulative distribution functions (cdf's).

Lemma 3.5.1. *Let F_n , $n \geq 0$, be cdf's. The following are equivalent:*

- (i) *For each $t \in \text{Disc}(F_0)^c$, $F_n(t_n) \rightarrow F_0(t)$ as $n \rightarrow \infty$ for some sequence $\{t_n : n \geq 1\}$ with $t_n \rightarrow t$.*
- (ii) *$F_n \Rightarrow F_0$; i.e., for each $t \in \text{Disc}(F_0)^c$, $F_n(t) \rightarrow F_0(t)$ as $n \rightarrow \infty$.*
- (iii) *For each $t \in \text{Disc}(F_0)^c$ and all sequences $\{t_n : n \geq 1\}$ with $t_n \rightarrow t$ as $n \rightarrow \infty$, $F_n(t_n) \rightarrow F_0(t)$ as $n \rightarrow \infty$.*

Proof. Clearly (iii) $\rightarrow$ (ii) $\rightarrow$ (i), so it suffices to show that (i) $\rightarrow$ (iii). Let $t \in \text{Disc}(F_0)^c$. Then, for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$F_0(t) - \epsilon \leq F_0(t - \delta) \leq F_0(t) \leq F_0(t + \delta) \leq F_0(t) + \epsilon. \quad (5.12)$$

Since F_0 is nondecreasing, it has at most countably many discontinuities. Let $t', t'' \in \text{Disc}(F_0)^c$ be such that $t - \delta < t' < t < t'' < t + \delta$. Given (i), there exist sequences $\{t'_n : n \geq 1\}$ and $\{t''_n : n \geq 1\}$ such that $t'_n \rightarrow t'$, $t''_n \rightarrow t''$, $F_n(t'_n) \rightarrow F_0(t')$ and $F_n(t''_n) \rightarrow F_0(t'')$ as $n \rightarrow \infty$. Let $\{t_n : n \geq 1\}$ be any sequence such that $t_n \rightarrow t$ as $n \rightarrow \infty$. Hence, there is an n_0 such that

$$F_0(t - \delta) < F_n(t'_n) \leq F_n(t_n) \leq F_n(t''_n) \leq F_0(t + \delta) \quad (5.13)$$

for all $n \geq n_0$. Combining (5.12) and (5.13), we see that

$$F_0(t) - \epsilon \leq F_n(t_n) \leq F_0(t) + \epsilon. \quad \blacksquare \quad (5.14)$$

Proof of Theorem 3.5.2. (a) Suppose that (5.5) holds. Then

$$F_t(x-) \equiv P(X(t) < x) \rightarrow P(L < x) \equiv F(x) \quad \text{as } t \rightarrow \infty \quad (5.15)$$

for each $x \in \text{Disc}(F)^c$. However,

$$\begin{aligned} F_t(x-) &= P(\phi_2(C(t)) < \phi_1(t) + x\psi(\phi_1(t))) \\ &= P(C(t) < \phi_2^{-1}(\phi_1(t) + x\psi(\phi_1(t)))) \end{aligned}$$

so that, by Lemma 3.4.1, $F_t(x) = P(S_{n(t)} > t)$ for any t such that

$$n(t) \equiv \phi_2^{-1}(\phi_1(t) + x\psi(\phi_1(t))) \quad (5.16)$$

is an integer. For such t ,

$$F_t(x-) = P(\psi(\phi_2(n(t)))^{-1}[\phi_1(S_{n(t)}) - \phi_2(n(t))] > -x(t)), \quad (5.17)$$

where

$$\begin{aligned} x(t) &= -[\phi_1(t) - \phi_2(n(t))]/\psi(\phi_2(n(t))) \\ &= x\psi(\phi_1(t))/\psi(\phi_1(t) + x\psi(\phi_1(t))) \rightarrow x \quad \text{as } t \rightarrow \infty \end{aligned} \quad (5.18)$$

by (5.16) and (5.4). Note that, for each positive integer n , we can find t_n such that $n(t_n) = n$, because ϕ_1 , ϕ_2 and ψ are nondecreasing and continuous, and $n(t) \rightarrow \infty$ as $t \rightarrow \infty$. Hence

$$G_n(x_n) \equiv P(\psi(\phi_2(n))^{-1}[\phi_1(S_n) - \phi_2(n)] < x_n) = F_{t_n}(x_n) \quad (5.19)$$

where $x_n = x(t_n) \rightarrow x$ as $n \rightarrow \infty$. Since $x \in \text{Disc}(F)^c$, $F_{t_n}(x_n) \rightarrow F(x)$ as $n \rightarrow \infty$. By Lemma 3.5.1, $G_n \Rightarrow F$, so that $Y(n) \Rightarrow -L$.

(b) Let the cdf G_n be defined by (5.19). Then

$$G_n(x) = P(S_n > \phi_1^{-1}(\phi_2(n) - x\psi(\phi_2(n)))) = P(A(t_n) < n)$$

for

$$t_n = \phi_1^{-1}(\phi_2(n) - x\psi(\phi_2(n))) \quad (5.20)$$

by Lemma 3.4.1. Thus, for F_t in (5.15) and t_n in (5.20), $F_{t_n}(x_n-) = G_n(x)$ for

$$\begin{aligned} x_n &= [\phi_2(n) - \phi_1(t_n)]/\psi(\phi_1(t_n)) \\ &= x\psi(\phi_2(n))/\psi(\phi_2(n) - x\psi(\phi_2(n))) \rightarrow x \quad \text{as } n \rightarrow \infty \end{aligned} \quad (5.21)$$

by (5.20) and (5.4). Hence, if $G_n(x) \rightarrow F(x)$ for $x \in \text{Disc}(F)^c$, then $F_{t_n}(x_n) \rightarrow F(x)$ as $n \rightarrow \infty$. By Lemma 3.5.1, $F_{t_n} \Rightarrow F$, so that $X(t_n) \Rightarrow L$.

(c) For any t , let n be such that $t_n \leq t < t_{n+1}$ for t_n in (5.20). Since $C(t_n) \leq C(t) \leq C(t_{n+1})$, it suffices to show that $C(t_n)$ and $C(t_{n+1})$ have the same limits with the normalization. It suffices to show that

$$\psi(\phi_1(t_{n+1}))^{-1}[\phi_2(C(t_n)) - \phi_1(t_{n+1})] \Rightarrow L, \quad (5.22)$$

which in turn holds if

$$\psi(\phi_1(t_{n+1}))/\psi(\phi_1(t_n)) \rightarrow 1 \quad (5.23)$$

and

$$[\phi_1(t_{n+1}) - \phi_1(t_n)]/\psi(\phi_1(t_n)) \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (5.24)$$

By (5.4) and (5.20), (5.23) is equivalent to (5.8). By (5.20) and (5.23), (5.24) is equivalent to (5.7): Applying (5.20) and dividing numerator and denominator by $\psi(\phi_2(n))$, we see that (5.21) becomes A_n/B_n , where

$$B_n = \psi(\phi_2(n) - x\psi(\phi_2(n)))/\psi(\phi_2(n)) \rightarrow 1 \quad \text{as } n \rightarrow \infty \quad (5.25)$$

by (5.4) and

$$A_n = [\phi_2(n+1) - x\psi(\phi_2(n+1)) - \phi_2(n) + x\psi(\phi_2(n))]/\psi(\phi_2(n)) \quad (5.26)$$

by (5.8) and (5.7). ■

Example 3.5.1. *It is possible that only subsequences converge.* To see that we can have $X(t_n) \Rightarrow L$ as $n \rightarrow \infty$ in Theorem 3.5.2 (b) without $X(t) \Rightarrow L$ as $t \rightarrow \infty$ for $X(t)$ in (5.5), let $\phi_1(t) = t$, $\phi_2(t) = t^2$ and $\psi(t) = 1$ for $t \geq 0$. Then (5.4) and (5.8) hold, but (5.7) does not:

$$\frac{\phi_2(n+1) - \phi_2(n)}{\psi(\phi_2(n))} = \frac{(n+1)^2 - n^2}{n^2} \rightarrow 0. \quad (5.27)$$

Let $P(L=0) = 1$ and let $S_n = n^2$, so that

$$\psi(\phi_2(n))^{-1}[\phi_1(S_n) - \phi_2(n)] = 0 = -L \quad \text{w.p.1 for all } n. \quad (5.28)$$

However, $C(t) = \sqrt{[t]}$ and $\phi_2(C(t)) = [t]$, so that

$$\psi(\phi_1(t))^{-1}[\phi_2(C(t)) - \phi_1(t)] = [t] - 1, \quad (5.29)$$

from which we see that

$$\varliminf_{t \rightarrow \infty} X(t) = -1 < 0 = \overline{\varlimsup}_{t \rightarrow \infty} X(t)$$

for $X(t)$ in (5.5). ■

A major theme here is obtaining probabilistic limits directly from deterministic limits. Thus it is natural to ask if there is a deterministic analog of Theorem 3.5.2 that implies Theorem 3.5.2. We show that there is. In particular, the following result implies parts (a) and (c) of Theorem 3.5.2.

Theorem 3.5.3. (deterministic analog of Theorem 3.5.2) *Let $\phi_1, \phi_2 \in \Lambda(\mathbb{R}_+)$ and let ψ be a continuous positive real-valued function on $[0, \infty)$ for which (5.4) holds*

(a) *If*

$$x(t) \equiv \psi(\phi_1(t))^{-1}[\phi_2(c(t)) - \phi_1(t)] \rightarrow \alpha \quad \text{in } \mathbb{R} \quad \text{as } t \rightarrow \infty, \quad (5.30)$$

then

$$y(n) \equiv \psi(\phi_2(n))^{-1}[\phi_1(s_n) - \phi_2(n)] \rightarrow -\alpha \quad \text{in } \mathbb{R} \quad \text{as } n \rightarrow \infty. \quad (5.31)$$

. (b) *If, in addition to (5.31) here, (5.7) and (5.8) above hold, then (5.30) here holds.*

Proof. (a) If (5.30) holds, then for all $\epsilon > 0$ there exists t_0 such that $\alpha - \epsilon \leq x(t) < \alpha + \epsilon$ for all $t \geq t_0$. Given that $x(t) < \alpha + \epsilon$,

$$\phi_2^{-1}(\phi_1(t) + (\alpha - \epsilon)\psi(\phi_1(t))) \leq c(t) < \phi_2^{-1}(\phi_1(t) + (\alpha + \epsilon)\psi(\phi_1(t))). \quad (5.32)$$

Let t be such that

$$n(t) \equiv \phi_2^{-1}(\phi_1(t) + (\alpha + \epsilon)\psi(\phi_1(t))) \quad (5.33)$$

is an integer. By Lemma 3.4.1,

$$s_{n(t)} > t. \quad (5.34)$$

Given (5.34),

$$y(n(t)) \equiv \psi(\phi_2(n(t)))^{-1}[\phi_1(s_{n(t)}) - \phi_2(n(t))] > -\alpha(t) \quad (5.35)$$

where

$$\alpha(t) \equiv \frac{\phi_1(t) - \phi_2(n(t))}{\psi(\phi_2(n(t)))} = \frac{-(\alpha + \epsilon)\psi(\phi_1(t))}{\psi(\phi_1(t) + (\alpha + \epsilon)\psi(\phi_1(t)))} \rightarrow -(\alpha + \epsilon) \quad (5.36)$$

by (5.4), so that $y(n(t)) > -(\alpha + 2\epsilon)$ for all $t \geq t_1 \geq t_0$. Since for each positive integer n , we can find t such that $n(t) = n$, there is an n_0 such that $y(n) > -(\alpha + 2\epsilon)$ for all $n \geq n_0$. Similarly, from the lower bound in (5.32),

we can conclude that $y(t) \leq -(\alpha - 2\epsilon)$ for all $n \geq n_1$. Since ϵ was arbitrary, the proof is complete.

(b) For any $\epsilon > 0$, there exists n_0 such that $-\alpha - \epsilon < y(n) \leq -\alpha + \epsilon$ for $n > n_0$. As a consequence,

$$\phi_1^{-1}(\phi_2(n) - (\alpha + \epsilon)\psi(\phi_2(n))) < s_n \leq \phi_1^{-1}(\phi_2(n) - (\alpha - \epsilon)\psi(\phi_2(n))) \quad (5.37)$$

for all $n \geq n_0$. Now let

$$t_n \equiv \phi_1^{-1}(\phi_2(n) - (\alpha - \epsilon)\psi(\phi_2(n))) . \quad (5.38)$$

By Lemma 3.4.1,

$$c(t_n) \geq n \quad (5.39)$$

and

$$\begin{aligned} x(t_n) &= \frac{\phi_2(c(t_n)) - \phi_1(t_n)}{\psi(\phi_1(t_n))} \\ &\geq \frac{\phi_2(n) - \phi_1(t_n)}{\psi(\phi_1(t_n))} \\ &= \frac{(\alpha - \epsilon)\psi(\phi_2(n))}{\psi(\phi_2(n) - (\alpha - \epsilon)\psi(\phi_2(n)))} \geq \alpha - 2\epsilon \end{aligned} \quad (5.40)$$

for $n \geq n_1 \geq n_0$ by (5.4). We now want to show that there is t_0 such that $x(t) \geq \alpha - 3\epsilon$ for all $t \geq t_0$. Consider t with $t_n \leq t < t_{n+1}$. Notice that

$$\phi_2(c(t_n)) - \phi_1(t_{n+1}) \leq \phi_2(c(t)) - \phi_1(t) \leq \phi_2(c(t_{n+1})) - \phi_1(t_n). \quad (5.41)$$

Since (5.41) holds, (5.40), (5.7) and (5.8) imply that there is t_0 such that $x(t) > \alpha - 3\epsilon$ for $t > t_0$. Similarly, using the lower bound in (5.37), we can deduce that for any $\epsilon > 0$ there exists t_0 such that $x(t) < \alpha + 3\epsilon$ for $t > t_0$. Since ϵ was arbitrary, the proof is complete. ■

3.6. Composition

We now turn to the composition map. We first state a preliminary lemma.

Lemma 3.6.1. *If $\phi \in \mathcal{R}(p)$ with $p > 0$, and $t^{-q}y(t) \rightarrow \mu > 0$, then*

$$\phi(y(t))/\phi(t^q) \rightarrow \mu^p \quad \text{as } t \rightarrow \infty. \quad (6.1)$$

Proof. For any $\epsilon > 0$, there is t_0 such that $t^q(\mu - \epsilon) < y(t) < t^q(\mu + \epsilon)$. Since ϕ is regularly varying with index p ,

$$\overline{\lim}_{t \rightarrow \infty} \frac{\phi(y(t))}{\phi(t^q)} \leq \lim_{t \rightarrow \infty} \frac{\phi(t^q(\mu + \epsilon))}{\phi(t^q)} \leq (\mu + \epsilon)^p$$

and

$$\underline{\lim}_{t \rightarrow \infty} \frac{\phi(y(t))}{\phi(t^q)} \geq \lim_{t \rightarrow \infty} \frac{\phi(t^q(\mu - \epsilon))}{\phi(t^q)} \geq (\mu - \epsilon)^p. \quad \blacksquare$$

Proposition 3.6.1. *Suppose that $\phi \in \mathcal{R}(p)$ with $p > 0$. If*

$$\phi(t)^{-1}X(t) \Rightarrow U \quad \text{and} \quad t^{-1}Y(t) \Rightarrow \mu \quad \text{in } \mathbb{R}, \quad (6.2)$$

then

$$\phi(t)^{-1}X(Y(t)) \Rightarrow \mu^p U \quad \text{in } \mathbb{R}. \quad (6.3)$$

Proof. Since the limit μ in (6.2) is deterministic, we have the joint limit

$$(\phi(t)^{-1}X(t), t^{-1}Y(t)) \Rightarrow (U, \mu) \quad \text{in } \mathbb{R}^2. \quad (6.4)$$

Use the Skorohod representation theorem to replace convergence in distribution in (6.4) with convergence w.p.1 (for special versions). By Lemma 3.6.1, $\phi(Y(t))/\phi(t) \rightarrow \mu^p$. Then

$$\frac{X(Y(t))}{\phi(t)} = \frac{\phi(Y(t))}{\phi(t)} \frac{X(Y(t))}{\phi(Y(t))} \rightarrow \mu^p U. \quad (6.5)$$

Finally, (6.5) implies (6.3). $\blacksquare$

Proposition 3.6.2. *Suppose that $\phi \in \mathcal{R}(p)$ with $0 < p < 1$. If*

$$\phi(t)^{-1}[X(t) - \lambda t, Y(t) - \mu t] \Rightarrow (U, V) \quad \text{in } \mathbb{R}^2 \quad (6.6)$$

then

$$\phi(t)^{-1}[X(Y(t)) - \lambda \mu t] \Rightarrow \mu^p U + \lambda V \quad \text{in } \mathbb{R}^2. \quad (6.7)$$

Proof. From (6.6) the regular variation condition, we have $t^{-1}Y(t) \Rightarrow \mu$ and $\phi(Y(t))/\phi(t) \rightarrow \mu^p$ as $t \rightarrow \infty$. Now replace convergence in distribution by convergence w.p.1 for special versions. Then

$$\frac{\phi(Y(t))}{\phi(t)} \frac{X(Y(t)) - \lambda Y(t)}{\phi(Y(t))} + \frac{\lambda Y(t) - \lambda \mu t}{\phi(t)} \rightarrow \mu^p U + \lambda V \text{ w.p.1 as } t \rightarrow \infty, \quad (6.8)$$

which implies (6.7). ■

There are two difficulties with Propositions (3.6.1) and (3.6.2) for applications. An obvious difficulty is that we may actually need the stronger conclusions giving limits in D in applications. The other difficulty is that it may be difficult to obtain the conditions. The joint limit in (6.6) holds if the component limits hold in $\mathbb{R}$ when $X(t)$ and $Y(t)$ are independent, but in most applications $X(t)$ and $Y(t)$ are actually dependent. A critical step then is to establish condition (6.6).

To illustrate, we may start with the sequence $\{(A_n, B_n) : n \geq 1\}$ of ordered pairs of nonnegative random variables. We may be able to determine that

$$\phi(n)^{-1}[A_n - n\lambda, B_n - n\mu^{-1}] \Rightarrow (U', V') \text{ as } n \rightarrow \infty \text{ in } \mathbb{R}^2 \quad (6.9)$$

and be interested in the asymptotic behavior of $A_{C(t)}$, where

$$C(t) = \max\{k \geq 1 : B_k \leq t\}, \quad t \geq 0. \quad (6.10)$$

From the second component of (6.9), we can determine that

$$\phi(t)^{-1}[C(t) - \mu t] \Rightarrow -\mu^{(1+p)} V' \quad (6.11)$$

from Theorem 3.5.1. However, we have difficulty directly expressing the joint limits of

$$\phi(n)^{-1}(A_n - n\lambda) \text{ as } n \rightarrow \infty \text{ and } \phi(t)^{-1}[C(t) - \mu t] \text{ as } t \rightarrow \infty. \quad (6.12)$$

The extension of (6.9) in D offers a resolution. We can hopefully extend (6.9) to

$$(\mathbf{A}_n, \mathbf{B}_n) \Rightarrow (\mathbf{U}, \mathbf{V}) \text{ in } D^2 \quad (6.13)$$

where

$$\begin{aligned} \mathbf{A}_n(t) &\equiv \phi(n)^{-1}[A_{[nt]} - \lambda nt] \\ \mathbf{B}_n(t) &\equiv \phi(n)^{-1}[B_{[nt]} - \mu^{-1}nt] \\ (\mathbf{U}(1), \mathbf{V}(1)) &\stackrel{d}{=} (U', V'). \end{aligned} \quad (6.14)$$

From (6.13), we can get

$$(\mathbf{A}_n, \mathbf{B}_n, \mathbf{C}_n) \Rightarrow (\mathbf{U}, \mathbf{V}, -\mu \mathbf{V} \circ \mu \mathbf{e}) \quad (6.15)$$

where

$$\mathbf{C}_n(t) \equiv \phi(n)^{-1}[C(nt) - \mu nt] . \quad (6.16)$$

We can then apply the composition map in D . In particular, letting

$$\Phi_n(t) \equiv n^{-1}C(nt) , \quad (6.17)$$

and

$$\mathbf{X}_n \equiv \phi(n)^{-1}[A_{C(nt)} - \lambda \mu nt] , \quad (6.18)$$

we obtain

$$\mathbf{X}_n = \mathbf{A}_n \circ \Phi_n + \mu \mathbf{C}_n \Rightarrow \mathbf{U} \circ \lambda \mathbf{e} + \mu \mathbf{C} \quad \text{in } D \quad (6.19)$$

under regularity conditions, by Theorem 13.3.2 in the book. As a consequence,

$$\phi(n)^{-1}[A_{C(n)} - \lambda \mu n] \Rightarrow \mathbf{U}(\lambda) + \mu \mathbf{C}(1) \quad \text{in } \mathbb{R}, \quad (6.20)$$

assuming that $P(1 \in \text{Disc}(\mathbf{U} \circ \lambda \mathbf{e} + \mu \mathbf{C})) = 0$.

Alternative approaches have been developed for dealing with this problem directly, starting with the Anscombe (1952, 1953) condition, see Gut (1988), but those conditions are essentially equivalent to $\mathbf{A}_n \Rightarrow \mathbf{U}$ with $P(\mathbf{U} \in C) = 1$.

3.7. Chapter Notes

The main results in this chapter are so basic that they no doubt have a long history, but we are unable to trace that history beyond our own work. Much related material, with emphasis on the classical case of partial sums of i.i.d. random variables, appears in Gut (1988).

We have primarily drawn upon Glynn and Whitt (1986, 1988) and Massey and Whitt (1994). Those papers contain further applications to queues related to the conservation law $L = \lambda W$. El-Taha and Stidham (1999) is closely related from that perspective. El-Taha and Stidham demonstrate the far-reaching implications possible from pointwise limits for single functions (sample-path analysis). Baccelli and Bremaud (1994) provide an alternative treatment of many of the same topics in the context of stationary processes. An overview of $L = \lambda W$ appears in Whitt (1991, 1992).

The strengthening of pointwise convergence to uniform convergence in Theorem 3.2.1 extends Theorem 4 of Glynn and Whitt (1988), which was in the form of Corollary 3.2.1. For the case $\phi(t) = t$, Proposition 3.3.1 is implication (iii) $\rightarrow$ (v) in Theorem 2(b) of Glynn and Whitt (1986). The more general version appears in Section 2.5 of El-Taha and Stidham (1999).

Theorem 3.5.1 here extends Theorem 4.2 of Massey and Whitt (1994) by allowing the space scaling function ψ to be regularly varying instead of a simple power. Lemma 3.5.1 is an improved statement of Lemma 4.1 of Massey and Whitt (1994). The deterministic basis for Theorem 3.5.2 in Theorem 3.5.3 is new here.

An extensive treatment of the composition map and convergence in distribution under a random time change appears in Gut (1988). The first few sections there provide useful perspective. A related result is the conservation law $Y = \lambda X$ in El-Taha and Stidham (1999).

Chapter 4

An Application to Simulation

4.1. Introduction

In Sections 5.9 and 10.4.4 of the book we showed how heavy-traffic stochastic-process limits for queues can be used to help plan queueing simulations. In this chapter we discuss another application of stochastic-process limits to simulation. We draw on Glynn and Whitt (1992a). In Section 4.2 we show how stochastic-process limits and the continuous-mapping approach can be used to determine general criteria for sequential stopping rules to be asymptotically valid.

Yet another application of stochastic-process limits and the continuous-mapping approach to simulation is contained in Glynn and Whitt (1992b). Glynn and Whitt (1992b) shows how stochastic-process limits and the continuous-mapping approach can be exploited to determine the asymptotic efficiency of simulation estimators. These two applications can be applied to queueing simulations, but they are not limited to queueing simulations.

4.2. Sequential Stopping Rules for Simulations

In this section, following Glynn and Whitt (1992a), we show how FCLTs and the continuous-mapping approach can be used to establish general conditions for the asymptotic validity of sequential stopping rules for stochastic simulations. The general conditions are expressed in terms of FCLTs and FWLLNs. The conditions allow the possibility of limit processes with discontinuous sample paths, but usually the limit process will be related to

Brownian motion, and thus have continuous sample paths. We use the composition and inverse maps to demonstrate the asymptotic validity.

The goal is to estimate a deterministic parameter $\alpha \in \mathbb{R}^k$. We start with an $\mathbb{R}^k$ -valued stochastic process $Y \equiv \{Y(t) : t > 0\}$ called the *estimation process*. We think of $Y(t)$ as being the estimate of α based on a simulation with runlength t . The results also apply to statistical estimation more generally, but we are especially concerned with simulation.

With simulation, a common problem is to estimate a steady-state mean vector α . The simulation may be used to generate a stochastic process $X \equiv \{X(t) : t \geq 0\}$, where $X(t) \Rightarrow X(\infty)$ in $\mathbb{R}^k$ as $t \rightarrow \infty$. We may then want to estimate the steady-state mean $\alpha \equiv EX(\infty) \equiv [EX^1(\infty), \dots, EX^k(\infty)]$ by the sample mean

$$Y(t) \equiv t^{-1} \int_0^t X(s) ds, \quad t > 0. \quad (2.1)$$

That is a common way for the estimation process Y to arise, but not the only way.

The simulator must select a runlength t . The runlength can be selected either in advance or sequentially while the simulation is in process. The principal disadvantage of selecting the runlength in advance is that the posterior precision of the estimator may not be appropriate. Since the volume of the confidence set (the width of a confidence interval in one dimension) is unknown in advance, the volume may be too large to be of practical use (meaning that the preassigned runlength was too small) or too small (meaning that computational resources were wasted in refining the estimator beyond the level of accuracy required).

We are interested in sequential procedures in which we let the simulation run until the volume of a confidence set achieves a prescribed value. That avoids the problems associated with preassigned runlengths, but new difficulties are introduced because the runlength is now randomly determined. The first difficulty is that we no longer have direct control of the amount of simulation time to be generated or the amount of computer time to be expended. Consequently, the runlength may turn out to be much longer than we want. On the other hand, it is possible that the runlength may turn out to be inappropriately short. This creates certain statistical difficulties that can compromise the accuracy of such procedures. For example, it is known that in many statistical settings, the point estimator and the variance estimator are positively correlated. Since the volume of a confidence set is typically determined by the variance estimator, this suggests that the confidence set volume will tend to be small when the point estimator is small.

Consequently, the resulting sequential procedure will tend to terminate early in situations in which the point estimator is too small, leading to possibly significant coverage problems for the confidence sets. Nevertheless, sequential stopping rules are of interest because of the possibility of automatically obtaining prescribed precision.

Various sequential stopping rules for simulation estimators have been proposed and investigated empirically. Among these are sequential procedures involving: batch means in Law and Carson (1979) and Law and Kelton (1982), regenerative simulation in Fishman (1977) and Lavenberg and Sauer (1977) and spectral methods in Heidelberger and Welch (1981a, b, 1983); see pages 81, 92, 97 and 103 of Bratley, Fox and Schrage (1987) for an overview. Unfortunately, however, the empirical evidence is not entirely encouraging. Evidently, care must be taken in the design and implementation of sequential procedures to avoid inappropriate early termination. On the positive side, the sequential procedures do tend to perform well when the run lengths are relatively long, which is achieved in part by having a suitably small prescribed volume for the confidence set. The observed good performance with small prescribed confidence set volumes is consistent with the asymptotic theory to be developed below. The asymptotic theory for general simulation estimators below is in turn consistent with the classical asymptotic theory associated with the sample mean of i.i.d. random variables; we cite references below.

4.2.1. The Mathematical Framework

To start, we assume that the estimation process Y satisfies a CLT, i.e.,

$$\phi(t)[Y(t) - \alpha] \Rightarrow \Gamma L \quad \text{in } \mathbb{R}^k \quad \text{as } t \rightarrow \infty, \quad (2.2)$$

where Γ is a nonsingular $k \times k$ scaling matrix and $\phi(t)$ is a real-valued scaling function with $\phi(t) \rightarrow \infty$ as $t \rightarrow \infty$. The common case for ϕ is $\phi(t) = t^{1/2}$, in which case the limit L in (2.2) typically is $N(0, I)$, a standard normal random vector with the identity matrix I as its covariance matrix, but we want to allow for other possibilities. With heavy-tailed probability distributions or long-range dependence, we might have $\phi(t) = t^\gamma$ for $\gamma < 1/2$ or, more generally, ϕ regularly varying with index γ . The treatment here generalizes Glynn and Whitt (1992a) by allowing regularly varying scaling functions instead of simple powers.

As a consequence of (2.2),

$$Y(t) \Rightarrow \alpha \quad \text{in } \mathbb{R}^k \quad \text{as } t \rightarrow \infty. \quad (2.3)$$

The limit (2.3) says that the estimation process is *weakly consistent*. Of course, weak consistency is a minimal requirement.

We assume that the confidence sets are all based on a bounded measurable subset A of $\mathbb{R}^k$ with $m(A) > 0$, where m is Lebesgue measure on $\mathbb{R}^k$. To obtain approximate $100(1 - \delta)\%$ confidence sets for α , we assume that

$$P(L \in A) = 1 - \delta \quad \text{and} \quad P(L \in \partial A) = 0, \quad (2.4)$$

where L is the limiting random variable in (2.2) and ∂A is the boundary of the set A , i.e., $\partial A = A^- - A^\circ$, where A^- and A° are the closure and interior of A . Given that we know A and Γ , we can let the *confidence set* be

$$\tilde{C}(t) \equiv Y(t) - \phi(t)\Gamma A, \quad (2.5)$$

where

$$z + QA \equiv \{x \in \mathbb{R}^d : \text{there exists } y \in A \text{ such that } x = z + Qy\}.$$

The confidence set $\tilde{C}(t)$ in (2.5) clearly depends on t . When the runlength t is specified in advance, the confidence set is asymptotically valid, in the sense of the following proposition.

Proposition 4.2.1. *If (2.2) and (2.4) hold, then*

$$P(\alpha \in \tilde{C}(t)) \rightarrow 1 - \delta \quad \text{as } t \rightarrow \infty$$

for $\tilde{C}(t)$ in (2.5).

Proof. Since Γ is nonsingular,

$$P(\alpha \in \tilde{C}(t)) = P(\Gamma^{-1}\phi(t)(Y(t) - \alpha) \in A),$$

but

$$\Gamma^{-1}\phi(t)(Y(t) - \alpha) \Rightarrow \Gamma^{-1}\Gamma L = L \quad \text{as } t \rightarrow \infty$$

by (2.2). Since (2.4) holds,

$$P(\Gamma^{-1}\phi(t)(Y(t) - \alpha) \in A) \rightarrow P(L \in A) = 1 - \delta \quad \text{as } t \rightarrow \infty$$

by Theorem 11.3.4 (v) in the book. ■

Of course, in applications the scaling matrix Γ is typically unknown, so that it too must be estimated. We assume that there is an estimator $\Gamma(t)$ that is weakly consistent, i.e.,

$$\Gamma(t) \Rightarrow \Gamma \quad \text{in } \mathbb{R}^{k^2} \quad \text{as } t \rightarrow \infty. \quad (2.6)$$

Given an estimator $\Gamma(t)$, $t > 0$, we can form approximate confidence sets based on $\Gamma(t)$. For that purpose, let

$$C(t) \equiv Y(t) - \phi(t)\Gamma(t)A . \quad (2.7)$$

We now extend Proposition 4.2.1 to include $\Gamma(t)$ instead of Γ .

Proposition 4.2.2. *If, in addition to the assumptions of Proposition 4.2.1 above, (2.6) holds, then*

$$P(\alpha \in C(t)) \rightarrow 1 - \delta \quad \text{as } t \rightarrow \infty$$

for $C(t)$ in (2.7).

Proof. By (2.2) above and Theorem 11.4.5 in the book,

$$(\Gamma(t), \phi(t)(Y(t) - \alpha)) \Rightarrow (\Gamma, \Gamma L) \quad \text{as } t \rightarrow \infty .$$

Then noting that matrix inversion is continuous at all nonsingular limits, we can deduce that $\Gamma(t)$ is nonsingular, and thus invertible, for all sufficiently large t and then apply the continuous mapping theorem to obtain

$$\Gamma(t)^{-1}\phi(t)(Y(t) - \alpha) \Rightarrow \Gamma^{-1}\Gamma L \quad \text{as } t \rightarrow \infty .$$

The rest of the proof is the same as the last part of the proof of Proposition 4.2.1. ■

We now use the confidence set $C(t)$ in (2.7) to define sequential stopping rules. Recall that, for a generic (measurable) set $B \subseteq \mathbb{R}^k$, $m(B)$ denotes the k -dimensional volume (Lebesgue measure) of the set. Of course, when $k = 1$ and B is an interval, $m(B)$ is just the length of the interval. We first consider the case in which the procedure terminates when the k^{th} root of the volume of the confidence region $C(t)$ drops below a prescribed level ϵ . [It is natural to use the k^{th} root, because $m(cB)^{1/k} = cm(B)^{1/k}$ for a scalar c .] We call such a procedure an *absolute-precision sequential stopping rule*. For such a rule, the time $\tilde{T}(\epsilon)$ at which the simulation terminates execution is defined by

$$\tilde{T}(\epsilon) = \inf\{t \geq 0 : m(C(t))^{1/k} < \epsilon\} . \quad (2.8)$$

Actually, this stopping rule needs to be modified, because $\tilde{T}(\epsilon)$ in (2.8) can terminate much too early if the estimator $\Gamma(t)$ is badly behaved for small t . To see this, suppose that $P(\Gamma(1) = 0, m(C(t)) = 1, 0 \leq t < 1) = 1$. In this case, $\tilde{T}(\epsilon) = 1$ for $\epsilon < 1$, so $C(\tilde{T}(\epsilon)) = Y(1)$ for $\epsilon < 1$. Hence, in this

example, $P(\alpha \in C(\tilde{T}(\epsilon))) = P(\alpha = Y(1))$ for $\epsilon < 1$. Hence convergence of the coverage probability of the region $C(T(\epsilon))$ to the nominal level $1 - \delta$ does *not* occur when we let $\epsilon \downarrow 0$.

In order for the asymptotic theory to be relevant to the sequential stopping problem, it is necessary that $T(\epsilon) \rightarrow \infty$ as $\epsilon \downarrow 0$. In other words, small values of the precision constant ϵ need to correspond to large values of simulation time. We can force the termination time to behave in this way if we inflate the volume $m(C(t))$ slightly. Let $a(t)$ be a strictly positive function that decreases monotonically to 0 as $t \rightarrow \infty$ and satisfies $a(t) = o(\phi(t))$ as $t \rightarrow \infty$, where ϕ is the scaling function in the CLT (2.2). Then set

$$T_1(\epsilon) \equiv \inf\{t \geq 0 : m(C(t))^{1/k} + a(t) < \epsilon\}. \quad (2.9)$$

Note that

$$T_1(\epsilon) \geq t_1(\epsilon) \equiv \inf\{t \geq 0 : a(t) < \epsilon\} \rightarrow \infty \quad \text{as } \epsilon \downarrow 0. \quad (2.10)$$

Thus the early termination associated with $\tilde{T}(\epsilon)$ in (2.8) is prevented by incorporating the deterministic function $a(t)$ in $T_1(\epsilon)$ in (2.9). For practical purposes, it remains to determine appropriate functions $a(t)$, though.

An alternative to the absolute-precision sequential stopping rule in (2.9) is a *relative-precision sequential stopping rule*. The basic idea here is that the simulation should terminate when the k^{th} root of the volume of the confidence region is less than an ϵ^{th} fraction of the norm of the parameter α , denoted by $\|\alpha\|$, under the additional condition that $\|\alpha\| > 0$. Since $Y(t)$ is an estimator for α , this suggests replacing $T_1(\epsilon)$ with

$$T_2(\epsilon) = \inf\{t \geq 0 : m(C(t))^{1/k} + \alpha(t) < \epsilon \|Y(t)\|\}. \quad (2.11)$$

The question now is: When are these sequentially stopping rules asymptotically valid? That is, when can we conclude that

$$P(\alpha \in C(T(\epsilon))) \rightarrow 1 - \delta \quad \text{as } \epsilon \downarrow 0 \quad (2.12)$$

for $T(\epsilon)$ being $T_1(\epsilon)$ in (2.9) or $T_2(\epsilon)$ in (2.11)?

It turns out that, unlike in Propositions 4.2.1 and 4.2.2, the assumed convergence in (2.2) and (2.6) is *not* enough to achieve asymptotic validity for the sequential stopping rules. That is for the same reason that CLTs involving random time change require extra conditions. However, we do obtain asymptotic validity if we replace the ordinary CLT in (2.2) by a FCLT and if we replace the ordinary WLLN in (2.6) by a SLLN or FWLLN.

(Recall that the SLLN implies a FSLLN by Corollary 3.2.1 in Chapter 3 here, which in turn implies a FWLLN, so that the SLLN is the stronger condition.)

For that purpose, we form scaled processes indexed by ϵ in the function space $D((0, \infty), \mathbb{R}^k)$. We work with time domain $(0, \infty)$ instead of $[0, \infty)$ in order to avoid having to deal with possible singularities in the estimation process Y at the origin $t = 0$. For example, such singularities occur in the special case in (2.1). Recall that $x_n \rightarrow x$ in $D((0, \infty), \mathbb{R}^d)$ if the restrictions converge in $D([t_0, t_1], \mathbb{R}^d)$ for all t_0, t_1 with $0 < t_0 < t_1 < \infty$.

Given the estimation process Y , the associated scaled estimation processes are

$$\mathbf{Y}_\epsilon(t) \equiv \phi(\epsilon^{-1})[Y(t/\epsilon) - \alpha], \quad t > 0. \quad (2.13)$$

For the results below we need to assume that the scaling function ϕ in (2.13) is regularly varying with index γ , denoted by $\phi \in \mathcal{R}(\gamma)$; see Appendix A in the book. We also assume that ϕ is a homeomorphism of $\mathbb{R}^+$, which implies that $\phi(0) = 0$ and ϕ is strictly increasing.

4.2.2. The Absolute-Precision Sequential Estimator

We first state a result for the absolute-precision sequential estimator $T_1(\epsilon)$ in (2.9).

Theorem 4.2.1. *Let $D \equiv D((0, \infty), \mathbb{R}^k)$ be endowed with the WM_2 or any other Skorohod topology. Suppose that*

$$\mathbf{Y}_\epsilon \Rightarrow \Gamma \mathbf{Z} \quad \text{in } D \quad \text{as } \epsilon \downarrow 0, \quad (2.14)$$

for $\mathbf{Y}_\epsilon$ in (2.13), where (2.4) holds with $L = \mathbf{Z}(1)$, $P(t \in \text{Disc}(\mathbf{Z})) = 0$ for all t , ϕ is a homeomorphism of $\mathbb{R}_+$, $\phi \in \mathcal{R}(\gamma)$ for $\gamma > 0$, and Γ is nonsingular. If, in addition,

$$\Gamma(t) \rightarrow \Gamma \quad \text{w.p.1 in } \mathbb{R}^{k^2} \quad \text{as } t \rightarrow \infty, \quad (2.15)$$

then as $t \rightarrow \infty$ or as $\epsilon \downarrow 0$

- (a) $\phi(t)[m(C(t))^{1/k} + a(t)] \rightarrow m(\Gamma A)^{1/k}$ w.p.1,
- (b) $\epsilon \phi(T_1(\epsilon)) \rightarrow m(\Gamma A)^{1/k}$ w.p.1,
- (c) $\epsilon^{-1} m(C(T_1(\epsilon)))^{1/k} \rightarrow 1$ w.p.1,
- (d) $\epsilon^{-1}[Y(T_1(\epsilon)) - \alpha] \Rightarrow m(\Gamma A)^{-1/k} \Gamma \mathbf{Z}(1)$ in $\mathbb{R}^k$,

(e) $P(\alpha \in C(T_1(\epsilon))) \rightarrow 1 - \delta$ (asymptotic validity).

In our proof of Theorem 4.2.1, we use the following lemma, which shows the scaling implications for the limit process $\mathbf{Z}$ from having the FCLT in (2.14) hold with the regularly varying scaling function ϕ in (2.13). The result is a consequence of Theorem 5.2.1 in the book, but we give a direct proof here.

Lemma 4.2.1. *If the FCLT (2.14) holds with $\phi \in \mathcal{R}(\gamma)$, $\gamma > 0$, for ϕ in (2.13), then*

$$\{\mathbf{Z}(ct) : t \geq 0\} \stackrel{d}{=} \{c^{-\gamma} \mathbf{Z}(t) : t \geq 0\} \quad (2.16)$$

for any $c > 0$.

Proof. Note that $\mathbf{Y}_\epsilon \circ c\epsilon \Rightarrow \mathbf{Z} \circ c\epsilon$ as $\epsilon \downarrow 0$. On the other hand,

$$\mathbf{Y}_\epsilon \circ c\epsilon = \frac{\phi(\epsilon^{-1})}{\phi(c\epsilon^{-1})} \mathbf{Y}_{\epsilon/c} \Rightarrow c^{-\gamma} \mathbf{Z} \quad \text{as } \epsilon \downarrow 0, \quad (2.17)$$

using the regular variation to get $\phi(\epsilon^{-1})/\phi(c\epsilon^{-1}) \rightarrow c^{-\gamma}$ as $\epsilon \downarrow 0$ for every $c > 0$; see Appendix A in the book. ■

Proof of Theorem 4.2.1. (a) Let

$$V(t) \equiv m(C(t))^{1/k} + a(t), \quad t > 0. \quad (2.18)$$

By the spatial invariance and scaling properties of Lebesgue measure m on $\mathbb{R}^k$,

$$\begin{aligned} m(C(t))^{1/k} &= m(Y(t) - \phi(t)^{-1} \Gamma(t) A)^{1/k} \\ &= m(-\phi(t)^{-1} \Gamma(t) A)^{1/k} = \phi(t)^{-1} m(\Gamma(t) A)^{1/k}. \end{aligned} \quad (2.19)$$

Since A is a bounded set, $\Gamma(t)A$ is contained in a bounded set for all sufficiently large t w.p.1. Thus, we can apply the bounded convergence theorem to deduce that

$$m(\Gamma(t)A)^{1/k} \rightarrow m(\Gamma A)^{1/k} \quad \text{w.p.1 as } t \rightarrow \infty. \quad (2.20)$$

Since Γ is nonsingular, $m(A) > 0$ and $a(t) = o(\phi(t)^{-1})$ as $t \rightarrow \infty$, (2.18) and (2.20) imply that

$$\phi(t)V(t) \rightarrow m(\Gamma A)^{1/k} > 0 \quad \text{w.p.1 as } t \rightarrow \infty. \quad (2.21)$$

(b) By the definition of $T_1(\epsilon)$ in (2.9), $V(T_1(\epsilon) - 1) \geq \epsilon$ and there exists a random variable $Z(\epsilon)$ with $0 \leq Z(\epsilon) \leq 1$ such that $V(T_1(\epsilon) + Z(\epsilon)) < \epsilon$. (Note that $V(t)$ is not necessarily monotone.) By (2.21) and the fact that $T_1(\epsilon) \rightarrow \infty$ w.p.1 as $\epsilon \downarrow 0$,

$$\limsup_{\epsilon \downarrow 0} \epsilon \phi(T_1(\epsilon)) \leq \limsup_{\epsilon \downarrow 0} \phi(T_1(\epsilon)) V(T_1(\epsilon) - 1) = m(\Gamma A)^{1/k} \quad \text{w.p.1} \quad (2.22)$$

and

$$\liminf_{\epsilon \downarrow 0} \epsilon \phi(T_1(\epsilon)) \geq \liminf_{\epsilon \downarrow 0} \phi(T_1(\epsilon)) (V(T_1(\epsilon)) + Z(\epsilon)) = m(\Gamma A)^{1/k} \quad \text{w.p.1.} \quad (2.23)$$

(c) Note that

$$m(C(T_1(\epsilon)))^{1/k} = \phi(T_1(\epsilon))^{-1} m(\Gamma(T_1(\epsilon))A)^{1/k} \quad (2.24)$$

and recall that $m(\Gamma(t)A) \rightarrow m(\Gamma A)$ w.p.1 as $t \rightarrow \infty$, so that $m(\Gamma(T_1(\epsilon))) \rightarrow m(\Gamma A)$ w.p.1 as $\epsilon \downarrow 0$. By (b), $\epsilon^{-1} \phi(T_1(\epsilon)) \rightarrow m(\Gamma A)^{-1/k}$. Hence

$$\begin{aligned} \epsilon^{-1} m(C(T_1(\epsilon)))^{1/k} &= \epsilon^{-1} \phi(T_1(\epsilon))^{-1} m(\Gamma(T_1(\epsilon))A)^{1/k} \\ &\rightarrow m(\Gamma A)^{-1/k} m(\Gamma A)^{1/k} = 1 \quad \text{w.p.1 as } \epsilon \downarrow 0 \end{aligned} \quad (2.25)$$

(d) From the assumed FCLT (2.14), $\mathbf{Z}_\epsilon \Rightarrow \Gamma \mathbf{Z}$ in $D((0, \infty), M_2)$ as $\epsilon \downarrow 0$, where

$$\mathbf{Z}_\epsilon(t) \equiv \mathbf{Y}_{1/\phi^{-1}(\epsilon^{-1})}(t) \equiv \epsilon^{-1} (Y(\phi^{-1}(\epsilon^{-1})t) - \alpha), \quad t > 0. \quad (2.26)$$

Now form the deterministic function

$$\psi_\epsilon(t) = \frac{\phi^{-1}(\epsilon^{-1}t)}{\phi^{-1}(\epsilon^{-1})}, \quad t > 0. \quad (2.27)$$

Since ϕ is a homeomorphism of $\mathbb{R}_+$, the inverse ϕ^{-1} exists and is itself an homeomorphism of $\mathbb{R}_+$. Moreover, since $\phi \in \mathcal{R}(\gamma)$, $\phi^{-1} \in \mathcal{R}(\gamma^{-1})$ by Theorem 1.5.12 of Bingham, Goldie and Tengels (1989). Hence

$$\psi_\epsilon \rightarrow \mathbf{e}^{1/\gamma} \quad \text{in } D \quad \text{as } \epsilon \downarrow 0. \quad (2.28)$$

We can apply the continuous-mapping theorem with the composition map taking $D \times D$ into D with (2.26)–(2.28), using Theorem 13.2.3 in the book, to conclude that

$$\mathbf{Z}'_\epsilon \Rightarrow \Gamma \mathbf{Z} \circ \mathbf{e}^{1/\gamma} \quad \text{in } (D, M_2) \quad \text{as } \epsilon \downarrow 0, \quad (2.29)$$

where

$$\mathbf{Z}'_\epsilon(t) \equiv (\mathbf{Z}_\epsilon \circ \psi_\epsilon)(t) \equiv \epsilon^{-1}Y(\phi^{-1}(\epsilon^{-1}t) - \alpha), \quad t > 0. \quad (2.30)$$

Finally, we can apply the continuous-mapping theorem with the composition map taking $D((0, \infty), \mathbb{R}^k) \times \mathbb{R}$ into $\mathbb{R}^k$ with (2.30), invoking Proposition 13.2.1 in the book and part (b) here, to obtain

$$\epsilon^{-1}Y(T_1(\epsilon) - \alpha) = \mathbf{Z}'_\epsilon(\epsilon\phi(T_1(\epsilon))) \Rightarrow \Gamma(\mathbf{Z} \circ \mathbf{e}^{1/\gamma})(m(\Gamma A)^{1/k}) \quad \text{in } \mathbb{R}^k, \quad (2.31)$$

where

$$(\mathbf{Z} \circ \mathbf{e}^{1/\gamma})(m(\Gamma A)^{1/k}) = \mathbf{Z}(m(\Gamma A)^{1/\gamma k}) \stackrel{d}{=} m(\Gamma A)^{-1/k} \mathbf{Z}(1) \quad (2.32)$$

by Lemma 4.2.1.

(e) Note that

$$\begin{aligned} P(\alpha \in C(T_1(\epsilon))) &= P(Y(T_1(\epsilon)) - \alpha \in \phi(T_1(\epsilon))^{-1}\Gamma(T_1(\epsilon))A) \\ &= P(\Gamma(T_1(\epsilon))^{-1}\phi(T_1(\epsilon))[Y_1(T_1(\epsilon)) - \alpha] \in A, \det(\Gamma(T_1(\epsilon))) \neq 0) \\ &\quad + P(Y(T_1(\epsilon)) - \alpha \in \phi(T_1(\epsilon))^{-1}\Gamma(T_1(\epsilon))A; \det(\Gamma(T_1(\epsilon))) = 0) \end{aligned} \quad (2.33)$$

Since $T_1(\epsilon) \rightarrow \infty$ w.p.1 and $\Gamma(t) \rightarrow \Gamma$ w.p.1, where Γ is nonsingular, $P(\det(\Gamma(T_1(\epsilon))) = 0) \rightarrow 0$ as $\epsilon \downarrow 0$, so that the second term on the right in (2.33) is negligible. On the other hand, for the first term,

$$\begin{aligned} \Gamma(T_1(\epsilon))^{-1}\phi(T_1(\epsilon))[Y(T_1(\epsilon)) - \alpha] &= \Gamma(T_1(\epsilon))^{-1}\epsilon\phi(T_1(\epsilon))\epsilon^{-1}[Y(T_1(\epsilon)) - \alpha] \\ &\Rightarrow \Gamma^{-1}m(\Gamma A)^{1/k}m(\Gamma A)^{-1/k}\Gamma\mathbf{Z}(1) = \mathbf{Z}(1) \end{aligned} \quad (2.34)$$

by parts (b) and (d). Hence, combining (2.33) and (2.34), we get

$$P(\alpha \in C(T_1(\epsilon))) \rightarrow P(\mathbf{Z}(1) \in A) = 1 - \delta, \quad (2.35)$$

because (2.4) holds with $L = \mathbf{Z}(1)$. ■

4.2.3. The Relative-Precision Sequential Estimator

We now state the analogous result for the relative-precision sequential estimator $T_2(\epsilon)$ in (2.11). Note that $T_2(\epsilon)$ behaves asymptotically like $T_1(\|\alpha\|\epsilon)$, as one would expect. In addition to the conditions in Theorem 4.2.2, we require that $Y(t) \rightarrow \alpha$ w.p.1 as $t \rightarrow \infty$. This is a reasonable condition, but it does not follow from the FCLT (2.14).

Theorem 4.2.2. *In addition to the conditions of Theorem 4.2.1, suppose that*

$$Y(t) \rightarrow \alpha \quad \text{w.p.1 in } \mathbb{R}^k \quad \text{as } t \rightarrow \infty,$$

where $\|\alpha\| > 0$. Then as $t \rightarrow \infty$ and $\epsilon \rightarrow 0$

$$(a) \quad \phi(t)[m(C(t))^{1/k} + a(t)]/\|Y(t)\| \rightarrow \|\alpha\|^{-1}m(\Gamma A)^{1/k} \quad \text{w.p.1},$$

$$(b) \quad \epsilon\phi(T_2(\epsilon)) \rightarrow \|\alpha\|^{-1}m(\Gamma A)^{1/k} \quad \text{w.p.1},$$

$$(c) \quad \epsilon^{-1}m(C(T_2(\epsilon)))^{1/k} \rightarrow \|\alpha\| \quad \text{w.p.1},$$

$$(d) \quad \epsilon^{-1}[Y(T_2(\epsilon)) - \alpha] \Rightarrow \|\alpha\|m(\Gamma A)^{-1/k}\Gamma Z(1) \quad \text{in } \mathbb{R}^k$$

$$(e) \quad P(\alpha \in C(T_2(\epsilon))) \rightarrow 1 - \delta \quad (\text{asymptotic validity}).$$

Since the proof of Theorem 4.2.2 closely parallels the proof of Theorem 4.2.1, we omit the proof of Theorem 4.2.2.

4.2.4. Analogs Based on a FWLLN

There are analogs of Theorems 4.2.1 and 4.2.2, where the SLLN for $\Gamma(t)$ in (2.15) is replaced by the weaker condition of a FWLLN. The w.p.1 limits in parts (a)–(c) of Theorems 4.2.1 and 4.2.2 are then replaced by FWLLNs and the CLT in (d) becomes a FCLT. Since the two results are similar, we only state the analog of Theorem 4.2.1.

Now we also generalize the framework by allowing a family of estimation processes indexed by ϵ . We start with processes $\{Y_\epsilon(t) : t \geq 0\}$ and $\{\Gamma_\epsilon(t) : t \geq 0\}$ for each $\epsilon > 0$. Then instead of (2.7), (2.9) and (2.13), let

$$\begin{aligned} C_\epsilon(t) &\equiv Y_\epsilon(t) - \phi(t)\Gamma_\epsilon(t)A, \\ T_{1\epsilon} &\equiv \inf\{t \geq 0 : m(C_\epsilon(t))^{1/k} + a(t) < \epsilon\}, \\ \mathbf{Y}_\epsilon(t) &= \phi(\epsilon^{-1})[Y_\epsilon(t/\epsilon) - \alpha], \quad t \geq 0. \end{aligned} \tag{2.36}$$

Then the limit will be for the following processes: For that purpose, we define the following random elements of D :

$$\begin{aligned} \Gamma_\epsilon(t) &\equiv \Gamma_\epsilon(t/\epsilon), \quad t > 0, \\ \mathbf{U}_\epsilon^1(t) &\equiv \phi(\epsilon^{-1})m(C_\epsilon(t/\epsilon))^{1/k}, \\ \mathbf{U}_\epsilon^2(t) &\equiv \epsilon T_{1\epsilon}(1/t\phi(\epsilon^{-1})), \\ \mathbf{U}_\epsilon^3(t) &\equiv \epsilon^{-1}m(C_\epsilon(T_{1\epsilon}(\epsilon/t)))^{1/k}, \\ \mathbf{Z}_\epsilon(t) &\equiv \epsilon^{-1}[Y_\epsilon(T_{1\epsilon}(\epsilon/t)) - \alpha]. \end{aligned} \tag{2.37}$$

Theorem 4.2.3. *Let the topology on D be one of WM_2 , SM_2 , WM_1 , SM_1 , WJ_1 or SJ_1 throughout. Suppose that the assumptions of Theorem 4.2.1 hold, except that (2.13) is replaced by (2.36) and condition (2.15) is replaced by*

$$\Gamma_\epsilon \Rightarrow \Gamma \mathbf{1} \quad \text{in } D^{k^2} \quad \text{as } \epsilon \downarrow 0, \quad (2.38)$$

for Γ_ϵ in (2.37) and $\mathbf{1}(t) = (1, \dots, 1)$ for all $t > 0$. Then

$$(\Gamma_\epsilon, \mathbf{U}_\epsilon^1, \mathbf{U}_\epsilon^2, \mathbf{U}_\epsilon^3, \mathbf{Z}_\epsilon) \Rightarrow (\Gamma \mathbf{1}, \mathbf{U}^1, \mathbf{U}^2, \mathbf{U}^3, \mathbf{Z}^1) \quad \text{in } D^{4+k} \quad \text{as } \epsilon \downarrow 0, \quad (2.39)$$

for $(\mathbf{U}_\epsilon^1, \mathbf{U}_\epsilon^2, \mathbf{U}_\epsilon^3, \mathbf{Z}_\epsilon)$ in (2.37), where

$$\begin{aligned} \mathbf{U}^1(t) &\equiv t^{-\gamma} m(\Gamma A)^{1/k}, & \mathbf{U}^2(t) &\equiv t^{1/\gamma} (\Gamma A)^{1/\gamma k} \\ \mathbf{U}^3(t) &= t^{-1} \quad \text{and} \quad \mathbf{Z}'(t) &\equiv m(\Gamma A)^{-1/k} \Gamma \mathbf{Z}(t^{1/\gamma}). \end{aligned} \quad (2.40)$$

Moreover,

$$P(\alpha \in C_\epsilon(T_{1\epsilon}(\epsilon))) \rightarrow 1 - \delta \quad (\text{asymptotic validity}). \quad (2.41)$$

In preparation for the proof of Theorem 4.2.3, we prove a lemma.

Lemma 4.2.2. *If $x_i \in D([a, b], \mathbb{R})$ for $i = 1, 2$, where $x_1(t), x_2(t) \geq c > 0$ for all t , then*

$$\|y_1 - y_2\| \leq c^{-2} \|x_1 - x_2\|$$

for $y_i(t) = 1/x_i(t)$, $a \leq t \leq b$.

Proof. Note that

$$|y_1(t) - y_2(t)| = \frac{|x_2(t) - x_1(t)|}{|x_1(t)| \cdot |x_2(t)|} \leq c^{-2} |x_2(t) - x_1(t)|. \quad \blacksquare$$

Corollary 4.2.1. *If $x_i \in D([a, b], \mathbb{R})$, $x_i(t) \geq c > 0$, and $y_i(t) = 1/x_i(t)$, $a \leq t \leq b$, $i = 1, 2$ then*

$$d(y_1, y_2) \leq (c^{-2} \vee 1) d(x_1, x_2)$$

where d is one of the J_1 , M_1 or M_2 metrics.

Proof. To illustrate, we do the J_1 case:

$$\begin{aligned} d(y_1, y_2) &= \inf_{\lambda \in \Lambda} \{ \|y_1 - y_2 \circ \lambda\| \vee \|\lambda - e\| \} \\ &\leq \inf_{\lambda \in \Lambda} \{ c^{-2} \|x_1 - x_2 \circ \lambda\| \vee \|\lambda - e\| \} \\ &\leq (c^{-2} \vee 1) d(x_1, x_2). \quad \blacksquare \end{aligned}$$

Proof of Theorem 4.2.3. First since the limit $\Gamma \mathbf{1}$ in (2.38) is deterministic, the two limits in (2.38) and (2.14) hold jointly (where $\mathbf{Y}_\epsilon$ is defined by (2.36) instead of (2.13)), by virtue of Theorem 11.4.5 in the book. Given those limits, we can apply the Skorohod representation theorem, as $\epsilon \downarrow 0$ through an arbitrary sequence, to replace the convergence in distribution by special versions converging w.p.1. Let the special versions be represented by the same notation. Since $\mathbf{1} \in C$, the convergence $\Gamma_\epsilon \rightarrow \Gamma \mathbf{1}$ in $D((0, \infty), \mathbb{R}^{d^2})$ is equivalent to uniform convergence over bounded intervals. Then, as in the proof of Theorem 4.2.1 (a), apply the bounded convergence theorem to get $m(\Gamma_\epsilon(t/\epsilon)A)^{1/k} \rightarrow m(\Gamma A)^{1/k}$ w.p.1 uniformly for $t \in [t_0, t_1]$ for any t_0, t_1 with $0 < t_0 < t_1 < \infty$. This yields w.p.1 convergence in $D((0, \infty), \mathbb{R})$ for the special versions. Since $\phi(t/\epsilon) \rightarrow 0$ as $\epsilon \downarrow 0$ uniformly in t for $t > t_0$, we obtain

$$\phi(t/\epsilon)V_\epsilon(t/\epsilon) \rightarrow m(\Gamma A)^{1/k} \quad \text{as } \epsilon \downarrow 0 \quad (2.42)$$

uniformly in $[t_0, t_1]$ for the special versions. Since $a(t/\epsilon) = o(\phi(t/\epsilon)^{-1})$, (2.42) implies that

$$\phi(t/\epsilon)m(C_\epsilon(t/\epsilon))^{1/k} \rightarrow m(\Gamma A)^{1/k} \quad \text{as } \epsilon \rightarrow 0 \quad (2.43)$$

uniformly in $[t_0, t_1]$. However, since $\phi \in \mathcal{R}(\gamma)$, $\phi(\epsilon^{-1})/\phi(t/\epsilon) \rightarrow t^{-\gamma}$ as $\epsilon \downarrow 0$ uniformly on $[t_0, t_1]$, by Theorem A.5 in Appendix A of the book. Thus,

$$\phi(\epsilon^{-1})V_\epsilon(t/\epsilon) \rightarrow t^{-\gamma}m(\Gamma A)^{1/k} \quad (2.44)$$

and

$$\phi(\epsilon^{-1})m(C_\epsilon(t/\epsilon))^{1/k} \rightarrow t^{-\gamma}m(\Gamma A)^{1/k} \quad \text{as } \epsilon \downarrow 0 \quad (2.45)$$

uniformly in $[t_0, t_1]$, again for the special versions, which implies the FCLT conclusion for $\mathbf{U}_\epsilon^1$ in $D((0, \infty), \mathbb{R})$. Turning to $\mathbf{U}_\epsilon^2$, we will show for the special versions that

$$\begin{aligned} \epsilon T_{1\epsilon}(1/t\phi(\epsilon^{-1})) &= \inf\{s \geq 0 : \phi(\epsilon^{-1})V_\epsilon(s/\epsilon) < t^{-1}\} \\ &= \inf\{s \geq 0 : \phi(\epsilon^{-1})^{-1}V_\epsilon(s/\epsilon)^{-1} > t\} \\ &\rightarrow \inf\{s \geq 0 : s^\gamma m(\Gamma A)^{-1/k} > t\} \\ &= t^{1/\gamma}m(\Gamma A)^{1/\gamma k}. \end{aligned} \quad (2.46)$$

uniformly in $[t_0, t_1]$. In the first line of (2.46), without loss of generality, we can replace $\phi(\epsilon^{-1})V_\epsilon(s/\epsilon)$ by $\max\{\phi(\epsilon^{-1})V_\epsilon(s/\epsilon), (2t_1)^{-1}\}$. Then we can invoke Corollary 4.2.1 above to show that the third line follows from (2.44). For $\mathbf{U}_\epsilon^3$, apply the continuous-mapping theorem with the composition map, using Theorem 13.2.3 in the book and (2.43) and (2.46) here, to get

$$\phi(\epsilon^{-1})m(C_\epsilon(T_{1\epsilon}(1/t\phi(\epsilon^{-1}))))^{1/k} \rightarrow t^{-1} \quad \text{as } \epsilon \rightarrow 0 \quad (2.47)$$

uniformly in $[t_0, t_1]$ or, equivalently,

$$\epsilon^{-1} m(C_\epsilon(T_{1\epsilon}(\epsilon/t)))^{1/k} \rightarrow t^{-1} \quad \text{as } \epsilon \rightarrow 0 \quad (2.48)$$

uniformly in $[t_0, t_1]$. Next, for $\mathbf{Z}_\epsilon$, apply the composition map again with the FCLT in (2.14) and the limit for $\mathbf{U}_\epsilon^2$ in (2.48) and part (b) to get

$$\mathbf{Z}_{\phi(\epsilon^{-1})^{-1}} \rightarrow \mathbf{Z}' \quad \text{in } D((0, \infty), \mathbb{R}^k) \quad \text{as } \epsilon \downarrow 0 \quad (2.49)$$

where

$$\mathbf{Z}'(t) \equiv \Gamma \mathbf{Z}(t^{1/\gamma} m(\Gamma A)^{1/\gamma k}) \stackrel{d}{=} m(\Gamma A)^{-1/k} \Gamma \mathbf{Z}(t^{1/\gamma})$$

and the topology is the same as for (2.14). Clearly, $\mathbf{Z}_\epsilon \rightarrow \mathbf{Z}'$ in $D((0, \infty), \mathbb{R}^k)$ as well. Finally, for (2.41), apply the projection map for $t = 1$ with the result $\mathbf{Z}_\epsilon \rightarrow \mathbf{Z}'$ just established. Then use the argument for Theorem 4.2.1 (e). ■

4.2.5. Examples

We conclude this section by giving several examples. We illustrate how the theorems can be applied by discussing a few specific estimation settings. These examples show that FCLT requirement for the estimation process Y in (2.14) is a mild hypothesis that is satisfied in virtually all practical contexts. However, some work may be required to establish the SLLN or FWLLN for the estimators $\Gamma(t)$ of the scaling matrix Γ . Our last example shows that we cannot instead use weak consistency of $\Gamma(t)$.

Example 4.2.1. (*Sample mean of IID random variables*). Suppose that α can be represented as $\alpha = EX$ for some real-valued r.v. X . For example, α might correspond to the expected number of customers served in a queue over the time interval $[0, T]$. Then α can be estimated by generating i.i.d. replicates $X_1, X_2, \dots$ of the r.v. X ; the resulting estimator for α is then the sample mean $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. The corresponding estimation process is $Y(t) = \bar{X}_{[t]}$, where $[t]$ is the greatest integer less than t and $\bar{X}_0 = 0$. If $EX^2 < \infty$, then Donsker's theorem, Theorem 4.3.2 in the book, asserts that the FCLT in (2.14) holds with $\phi(\epsilon^{-1}) = \epsilon^{-1/2}$ in (2.13), $\Gamma = \sigma$, where $\sigma^2 = \text{var } X$, and $\mathbf{Z}(t) = \mathbf{B}(t)/t$, where $\mathbf{B}$ is Brownian motion. Note that $\mathbf{Z}(1) = {}_d N(0, 1)$. The typical choice for the set A in this setting is the interval $[-z(\delta), z(\delta)]$, where $z(\delta)$ is chosen to satisfy $P(N(0, 1) \leq z(\delta)) = 1 - \delta/2$. Of course, it is well known that

$$\Gamma_n \equiv \left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \right]^{1/2} \rightarrow \sigma \quad \text{w.p.1 as } n \rightarrow \infty. \quad (2.50)$$

Suppose that $\sigma^2 > 0$. Setting $\Gamma(t) = \Gamma_{[t]}$, we have the strong consistency required by Theorems 4.2.1 and 4.2.2. Hence both the absolute-precision and relative-precision stopping rules $T_1(\epsilon)$ and $T_2(\epsilon)$ are asymptotically valid for this example when the precision-constant ϵ shrinks to 0. In this setting, Theorems 4.2.1 and 4.2.2 reproduce the classical results of Chow and Robbins (1965), Starr (1966) and Nadas (1969); see Chapter 7 of Siegmund (1985) and Section 8.8 of Wetherill and Glazebrook (1986). (See Anscombe (1952, 1953) for related earlier work.) Implementation considerations are discussed in Law, Kelton and Koenig (1981).

Example 4.2.2. (*The sample mean of IID random vectors*). Now we consider the case in which α can be represented as $\alpha = EX$, where X is $\mathbb{R}^k$ -valued. Assume that $E\|X\|^2 < \infty$. As in Example 4.2.1, we can estimate α via the sample mean $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$, where X_i 's are i.i.d. copies of X . Setting $Y(t) = \bar{X}_{[t]}$, we obtain the FCLT (2.14) from the k -dimensional version of Donsker's theorem, Theorem 4.3.5 in the book, where now $\mathbf{Z}(t) = \mathbf{B}(t)/t$, $\mathbf{B}$ is k -dimensional standard Brownian motion (composed of k independent one-dimensional standard Brownian motions) and $\Gamma\Gamma^t$ is the covariance matrix C of X . We assume that C is positive definite. Note that $\mathbf{Z}(1) = \mathbf{B}(1) = {}_d N(0, I)$, where I is the identity matrix. In this k -dimensional setting, we can assume that A is the k -sphere $\{x : \|x\| \leq w(\delta)\}$, where $w(\delta)$ is chosen so that

$$P\{\|N(0, I)\|^2 \leq w^2(\delta)\} = P\{\mathcal{X}_k^2 \leq w^2(\delta)\} = 1 - \delta, \quad (2.51)$$

with $\mathcal{X}_k^2$ being a chi-squared r.v. with k degrees of freedom. Let

$$C_n = \frac{1}{n} \sum_{i=1}^n X_i X_i^t - \bar{X}_n \bar{X}_n^t \quad (2.52)$$

(writing all k -vectors as column vectors). Then $C_n \rightarrow C$ a.s. as $n \rightarrow \infty$. Let Γ_n be obtained by taking the Cholesky factorization of C_n , so that Γ_n is a lower triangular matrix such that $C_n = \Gamma_n \Gamma_n^t$; see pages 164 and 165 of Bratley, Fox and Schrage (1987). It follows that $\Gamma_n \rightarrow \Gamma$ w.p.1 as $n \rightarrow \infty$, since Cholesky factors are continuous at positive definite matrices. Setting $\Gamma(t) = \Gamma_{[t]}$, we again have the strong consistency required by Theorems 4.2.1 and 4.2.2. Thus we have proved that the absolute-precision and relative-precision stopping rules $T_1(\epsilon)$ and $T_2(\epsilon)$ are asymptotically valid for sequential stopping of multiple performance measure stochastic simulations. In this setting, Theorems 4.2.1 and 4.2.2 reproduce results by Gleser (1965), Albert (1966) and Srivastava (1967); see Section 5.5 of Govindarajulu (1987).

Example 4.2.3. (*Functions of sample means*). Let X be an $\mathbb{R}^k$ -valued random vector and let $\mu = EX$. Suppose that α can be represented as $\alpha = g(\mu)$ for some (known) real-valued function $g : \mathbb{R}^k \rightarrow \mathbb{R}$. An example of this occurs in the ratio estimation setting, in which $k = 2$ and $g(x, y) = x/y$. Because the steady state of a regenerative stochastic process can be expressed as a ratio of two means, this estimation setting subsumes that of regenerative steady-state simulation. Of course, this observation lies at the heart of the regenerative method of steady-state simulation; see, for example, Crane and Lemoine (1977).

In this nonlinear setting, we estimate α via $Y(t) = g(\bar{X}_{[t]})$, where X_i are i.i.d. random vectors as in Example 4.2.2. Suppose that $E\|X\|^2 < \infty$ and that g is continuously differentiable in a neighborhood of μ . In addition, we require that $\nabla g(\mu) \neq 0$ and that the covariance matrix C of X is positive definite. Then Theorem 3 of Glynn and Whitt (1992b) implies that the FCLT in (2.14) holds with $\phi(\epsilon^{-1}) = \epsilon^{-1/2}$, $\mathbf{Z}(t) = \mathbf{B}(t)/t$ and $\Gamma = \sigma$ as in Example 4.2.1, but with

$$\sigma = (\nabla g(\mu)^t C \nabla g(\mu))^{1/2} .$$

Let C_n be defined as in Example 4.2.2 and note that

$$[\nabla g(Y(t))^t C_{[t]} \nabla g(Y(t))]^{1/2} \rightarrow \sigma \text{ w.p.1 as } t \rightarrow \infty .$$

Hence we have the strong consistency required for the application of Theorems 4.2.1 and 4.2.2. As a consequence, we are assured that the stopping rules $T_1(\epsilon)$ and $T_2(\epsilon)$ are again asymptotically valid in this estimation setting. In particular, in the regenerative simulation setting, we recover the asymptotic theory developed by Lavenberg and Sauer (1977).

Example 4.2.4. (*The jackknife*). Consider the estimation problem of Example 4.2.3 in which our goal is to estimate $\alpha = g(\mu)$, where μ can be expressed as $\mu = EX$ and g is real-valued. One practical difficulty with the estimator suggested in Example 4.2.3 is that it tends to be significantly affected by bias problems induced by the presence of the nonlinearity in g . One way to address the small-sample bias problem that this nonlinearity creates is to jackknife the estimator. Specifically, let $\alpha(n) = g(\bar{X}_n)$ and, for $1 \leq i \leq n$, let

$$\begin{aligned} \bar{X}_{in} &= \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^n X_j, & \alpha_i(n) &= g(\bar{X}_{in}) , \\ \tilde{\alpha}_i(n) &= n\alpha(n) - (n-1)\alpha_i(n) . \end{aligned} \tag{2.53}$$

Then the estimator $Y_n = n^{-1} \sum_{i=1}^n \tilde{\alpha}_i(n)$ is the *jackknife estimator* of α . Let $Y(t) = Y_{[t]}$. It is shown in Glynn and Heidelberger (1989) that if $E\|X\|^3 < \infty$ and g is twice continuously differentiable in a neighborhood of μ , then the FCLT in (2.14) holds where σ and $Z(t)$ are as in Example 4.2.3. Since the form of the FCLT is the same as for Example 4.2.3, the jackknife has the same asymptotic efficiency as the estimator of Example 4.2.3. However, as argued in Miller (1964, 1974), the jackknife estimator typically possesses superior small-sample bias properties.

Two estimators for the scaling constant $\sigma = [\nabla g(\mu)^t C \nabla g(\mu)]^{1/2}$ are possible. One approach is to use the estimator $\sigma(t) = [\nabla g(Y(t))^t C_{[t]} \nabla g(Y(t))]^{1/2}$ suggested in Example 4.2.3. Theorem 4(i) of Glynn and Heidelberger (1989) shows that $Y(t) \rightarrow \alpha$ w.p.1 as $t \rightarrow \infty$, under the conditions stated here. Since $C_n \rightarrow C$ w.p.1, it follows that $\sigma(t) \rightarrow \sigma$ w.p.1 as $t \rightarrow \infty$. Hence sequential stopping procedures based on the jackknife point estimator and the “variance” estimator $\sigma^2(t)$ are asymptotically valid by Theorems 4.2.1 and 4.2.2, provided that $\sigma^2 > 0$.

An alternative estimator for the scaling constant σ is given by the jackknife variance estimator $\sigma_J(t)$:

$$\sigma_J(t) = \left(\frac{1}{[t]} \sum_{i=1}^{[t]} (\tilde{\alpha}_i([t]) - Y(t))^2 \right)^{1/2}. \quad (2.54)$$

Although it is known that $\sigma_J^2(t) \Rightarrow \sigma^2$ as $t \rightarrow \infty$ under suitable regularity conditions, we need convergence w.p.1 in order to satisfy the hypothesis of Theorems 4.2.1 and 4.2.3. However, Theorem 4 of Glynn and Whitt (1992a) establishes the following result.

Theorem 4.2.4. *If g is continuously differentiable in a neighborhood of μ and $E|X|^2 < \infty$, then*

$$\sigma_J^2(t) \rightarrow \sigma^2 = \nabla g(\mu)^t C \nabla g(\mu) \text{ w.p.1 as } t \rightarrow \infty \quad (2.55)$$

for $\sigma_J^2(t)$ in (2.54). Thus the sequential stopping rules $T_1(\epsilon)$ and $T_2(\epsilon)$ may be applied to jackknife point estimators in conjunction with the jackknifed variance estimator $\sigma_J^2(t)$.

Example 4.2.5. (*A steady-state mean*). Suppose that our goal is to estimate the steady-state mean vector α of an $\mathbb{R}^k$ -valued stochastic process $X = \{X(t) : t \geq 0\}$. We assume that X satisfies an FCLT, namely,

$$\mathbf{X}_\epsilon \Rightarrow \Gamma \mathbf{B} \quad \text{in } D((0, \infty), \mathbb{R}^k) \quad \text{as } \epsilon \downarrow 0 \quad (2.56)$$

where

$$\mathbf{X}_\epsilon(t) \equiv \epsilon^{-1} \left(\int_0^{t/\epsilon} X(s) ds - t\alpha \right), \quad t > 0. \quad (2.57)$$

and $\mathbf{B}$ is a standard $\mathbb{R}^k$ -valued Brownian motion. It is easily shown that (2.56) implies that

$$Y(t) \equiv t^{-1} \int_0^t X(s) ds \Rightarrow \alpha \quad \text{as } t \rightarrow \infty. \quad (2.58)$$

Hence (2.56) implies that the centering vector α appearing there is indeed the steady-state mean of X . Another easy consequence of (2.56) is that the FCLT (2.14) holds with $\phi(\epsilon^{-1}) = \epsilon^{-1/2}$ and $\mathbf{Z}(t) = \mathbf{B}(t)/t$.

It turns out that (2.56) is typically satisfied for most “real-world” steady-state simulations. In particular, a great variety of different assumptions on the structure of the process X give rise to FCLTs of the form (2.56); see Section 4.4 in the book and Section 2.3 here.

The primary difficulty in applying Theorems 4.2.1–4.2.3 arises in the construction of a process $\Gamma(t)$ such that $\Gamma(t) \rightarrow \Gamma$ w.p.1 as $t \rightarrow \infty$ or $\Gamma_\epsilon \Rightarrow \Gamma \mathbf{1}$ in $D(0, \infty)$ as $\epsilon \downarrow 0$. Since $\Gamma \Gamma^t$ is the covariance matrix of the limiting Brownian motion, this is equivalent to the construction of a strongly consistent estimator $C(t)$ for the *time-average covariance matrix* $C = \Gamma \Gamma^t$ of X . In general, this is known to be a challenging problem.

Suppose that X is regenerative, with regeneration times $0 = \tau_0 < \tau_1 < \tau_2 < \dots$. Suppose that $E(\int_{\tau_1}^{\tau_2} |X(s) - \alpha|^2 ds) < \infty$ and that $E(\tau_2 - \tau_1) < \infty$. Let $N(t) = \max\{n \geq 0 : \tau_n \leq t\}$. Then it is easily proved that

$$C(t) = \frac{1}{t} \sum_{i=1}^{N(t)} \int_{\tau_{i-1}}^{\tau_i} [X(s) - Y(t)][X(s) - Y(t)]^t ds \quad (2.59)$$

is strongly consistent for C , where $C = \Gamma \Gamma^t$ and Γ is the scaling matrix appearing in (2.56). Thus when X is regenerative, the sequential stopping rules $T_1(\epsilon)$ and $T_2(\epsilon)$ are asymptotically valid. Of course, when X is scalar, we already established this result in Example 4.2.3.

For nonregenerative processes, less is known about the strong consistency of estimators $C(t)$ for the steady-state covariance matrix. However, Glynn and Iglehart (1988) and Damerdjani (1991, 1994) have recently used strong approximation techniques to establish strong consistency for a broad class of estimators for C . Thus Theorems 4.2.1 and 4.2.2 prove that these estimators do indeed lead to asymptotically valid sequential procedures.

Our theory for this example provides theoretical support complementing previous work by Fishman (1977), Law and Carson (1979) and Law and Kelton (1982). They develop specific empirically based sequential stopping rules for steady-state simulations.

Example 4.2.6. (*Kiefer-Wolfowitz stochastic approximation*). This example is interesting, in part, because it illustrates that the FCLT (2.14) can hold for the estimator with a subcanonical convergence rate; in particular, here $\phi(\epsilon^{-1}) = \epsilon^{-1/3}$. For other examples of noncanonical estimator convergence rates, see Fox and Glynn (1989) and Sections 5 and 6 of Glynn and Whitt (1992b). Suppose that we are given a real-valued smooth function $\beta(\theta)$, which can be represented as $\beta(\theta) = EZ(\theta)$. Assume that our goal is to compute the parameter $\alpha \equiv \theta^*$ minimizing β . If θ is scalar, we can apply the following Kiefer-Wolfowitz stochastic approximation algorithm:

$$\theta_{n+1} = \theta_n - c_n X_{n+1}, \quad (2.60)$$

where $\{c_n : n \geq 0\}$ is a sequence of (deterministic) nonnegative constants,

$$\begin{aligned} P(X_{n+1} \in A | \theta_0, X_0, \dots, \theta_n, X_n) = \\ P([Z(\theta_0 + h_{n+1}) - Z(\theta_0 - h_{n+1})]/2h_{n+1} \in A), \end{aligned} \quad (2.61)$$

$Z(\theta_0 + h_{n+1})$ and $Z(\theta_0 - h_{n+1})$ are generated independently of one another and $\{h_n : n \geq 1\}$ is another sequence of deterministic constants. Suppose that $c_n = cn^{-1}$ and $h_n = hn^{-1/3}$, $c, h > 0$. Let $Y(t) = \theta_{[t]}$. For this problem, Ruppert (1982) showed that under suitable regularity conditions, the FCLT in (2.14) holds for $\mathbf{Y}_\epsilon$ in (2.13) with $\phi(\epsilon^{-1}) = \epsilon^{-1/3}$, $\Gamma = \kappa$, $\mathbf{Z}(t) = t^{-b}\mathbf{B}(t^{2\eta+1})$, $\mathbf{B}$ is a standard Brownian motion, $b = c\beta''(\theta^*)$, $\eta = b - 5/6$, $\kappa^2 = c^2\sigma^2/(2\eta + 1)(4h^2)$ and $\sigma^2 = 2\text{var } \mathbf{Z}(\theta^*)$.

The construction of a strongly consistent estimator for $\Gamma \equiv \kappa$ involves more work. For some directions on how to obtain such an estimator, see page 189 of Venter (1967).

Example 4.2.7. (*Robbins-Monro stochastic approximation*). As in Example 4.2.6, suppose that our goal is to estimate the minimizer θ^* of a smooth function $\beta: \mathbb{R} \rightarrow \mathbb{R}$. However, we assume here that we can represent the derivative β' as an expectation; that is, there exists a process $Z(\theta)$ such that $\beta'(\theta) = EZ(\theta)$. [In Example 4.2.6 we assumed only that the function values $\beta(\theta)$ could be represented as expectations.] To calculate θ^* in this setting, we can use the Robbins-Monro stochastic approximation algorithm,

which is based on (2.60), where $\{c_n : n \geq 0\}$ is sequence of (deterministic) nonnegative constants and

$$P(X_{n+1} \in A | \theta_0, X_0, \dots, \theta_0, X_n) = P(Z(\theta_n) \in A). \quad (2.62)$$

Suppose that our estimator is $Y(t) = \theta_{[t]}$ and $c_n = cn^{-1}$ with $c > 0$. Then Ruppert (1982) showed that under suitable regularity hypotheses, the FCLT in (2.14) holds for $\mathbf{Y}_\epsilon$ in (2.13) with $\phi(\epsilon^{-1}) = \epsilon^{-1/2}$, $\Gamma = \kappa$, $\mathbf{Z}(t) = t^{-(D+1)}\mathbf{B}(t^{2D+1})$, $D = c\beta'(\theta^*) - 1$, $\kappa^2 = c^2\sigma^2(2D+1)^{-1}$, $\sigma^2 = \text{var } \mathbf{Z}(\theta^*)$ and $\mathbf{B}$ is a standard Brownian motion.

Construction of a strongly consistent estimator for $\Gamma \equiv \kappa$ follows from results established by Venter (1967). When this estimator is used, the sequential stopping rule $T_1(\epsilon)$ reduces to one studied by McLeish (1976).

Example 4.2.8. (*The Hill estimator*). The framework of Theorems 4.2.1–4.2.3 has been made quite general, so that there can be many applications. One intended application is to estimation problems associated with heavy-tailed probability distributions and long-range dependence. If we use the direct (naive) estimators, e.g., the time average for the steady-state mean, then we anticipate that the FCLT in (2.14) will typically hold with ϕ in (2.13) satisfying $\phi(\epsilon^{-1})/\epsilon^{-1/2} \rightarrow 0$ as $\epsilon \downarrow 0$. A common case would be $\phi(\epsilon^{-1}) = \epsilon^{-\gamma}$ or $\phi \in \mathcal{R}(\gamma)$ for $0 < \gamma < 1/2$. A major new difficulty, however, is that now the scaling exponent γ is typically unknown.

Thus, attention naturally shifts to estimating the scaling parameter γ . Estimating the parameter γ is challenging even from observations of i.i.d. random variables. One approach is via the Hill estimator. Recent results of Resnick and Stărică (1997) show that Theorems 4.2.1–4.2.3 can be applied.

The setting is a sequence $\{X_n : n \geq 1\}$ of i.i.d. positive random variables having cdf F , where $F^c \equiv 1 - F \in \mathcal{R}(-\alpha)$ for $\alpha > 0$, i.e.

$$F^c(tx)/F^c(t) \rightarrow x^{-\alpha} \quad \text{as } t \rightarrow \infty. \quad (2.63)$$

The goal is to estimate the tail index α . For n given, let $X_{(i)}$ be the i^{th} largest among the first n . The Hill estimator based on the k upper order statistics is

$$H_{k,n} = k^{-1} \sum_{i=1}^k \log \left(\frac{X_{(i)}}{X_{(k+1)}} \right). \quad (2.64)$$

The Hill estimator is known to be consistent if $k \equiv k(n)$ satisfies $k(n) \rightarrow \infty$ and $k(n)/n \rightarrow 0$ as $n \rightarrow \infty$. Given a specific function $k(n)$, the Hill estimator is a single sequence of random variables $\{H_{k(n),n} : n \geq 1\}$. Resnick and

Stărică (1997) show that the Hill estimator also satisfies a FCLT. To state it, let

$$\mathbf{Y}_n(t) \equiv H_{[k_n t], n}, \quad t \geq 0, \quad (2.65)$$

where $[x]$ is the least integer greater than or equal to x . The FCLT states that, under regularity conditions, including $k(n) \rightarrow \infty$ and $k(n)/n \rightarrow 0$,

$$k(n)[\mathbf{Y}_n - \alpha^{-1}\mathbf{1}] \Rightarrow \alpha^{-1}\mathbf{Z} \quad \text{in } D \quad \text{as } n \rightarrow \infty, \quad (2.66)$$

with $\mathbf{Z}(t) = t^{-1}\mathbf{B}(t)$, where $\mathbf{B}$ is standard Brownian motion. Notice that here we use the more general framework in (2.36) in which there is a family of estimation processes indexed by $\epsilon > 0$. (Here we have used $n \rightarrow \infty$ instead of $\epsilon \downarrow 0$.)

Also notice that in this special case the scaling matrix Γ in (2.14) is just α^{-1} . So, with $\mathbf{Y}_n$ in (2.65), we estimate α and Γ simultaneously. As a consequence of the FCLT in (2.66), we have the associated FWLLN

$$\mathbf{Y}_n \Rightarrow \alpha^{-1}\mathbf{1} \quad \text{in } D \quad \text{as } n \rightarrow \infty \quad (2.67)$$

needed in Theorem 4.2.3. It is also known that $Y_n(t) \rightarrow \alpha^{-1}$ w.p.1 as $n \rightarrow \infty$ under regularity conditions.

Given the FCLT in (2.66) and the FWLLN in (2.67), the conditions of Theorem 4.2.3 are satisfied. Hence sequential stopping rules are asymptotically valid for the Hill estimator too.

Example 4.2.9. (*Sample mean with infinite variance*). One can also estimate a mean by the sample mean of i.i.d. random variables when the random variables X_i have finite mean but infinite variance. As in Example 4.2.1, the estimation process can be $Y(t) = \bar{X}_{[t]}$, where $\bar{X}_0 = 0$, although it is often better to use alternative robust estimators such as trimmed means or to estimate other quantities such as the median. Under regularity conditions, FCLT (2.14) is valid with $\phi \in \mathcal{R}(1 - \alpha^{-1})$ for some α , $1 < \alpha < 2$, where ϕ is the scaling function in (2.13). The topology on D can be the J_1 topology. The limit process $\mathbf{Z}(t)$ is then $t^{-1}\mathbf{S}_\alpha(t)$, where $\{\mathbf{S}_\alpha(t) : t \geq 0\}$ is a stable process of index α , which depends on two parameters in addition to α : a scale parameter σ and a skewness parameter β , $-1 \leq \beta \leq 1$. Unfortunately, in order to form confidence sets we need to estimate the scaling function ϕ and the parameters σ and β .

Suppose that we consider the special case in which X_i is nonnegative and is assumed to have an asymptotic power tail, i.e.

$$F^c(t) \equiv P(X > t) \sim At^{-\alpha} \quad \text{as } t \rightarrow \infty \quad (2.68)$$

for positive constants A and α , $1 < \alpha < 2$. Under condition (2.68), the FCLT (2.14) holds with $\phi(\epsilon^{-1}) = \epsilon^{-(1-\alpha^{-1})}$ and limit process $\Gamma Z(t)$ where $Z(t)$ is a stable process with index α , scale $\sigma = 1$ and skewness 1. Hence, in this special case it suffices to estimate only the two parameters α and Γ .

Suppose that $\hat{\alpha}_\epsilon$ is an estimate of α with the property that

$$(\hat{\alpha}_\epsilon^{-1} - \alpha^{-1}) \log(\epsilon^{-1}) \rightarrow 0 \quad \text{w.p.1} \quad \text{as} \quad \epsilon \downarrow 0. \quad (2.69)$$

Given (2.69),

$$\hat{\phi}(\epsilon^{-1})/\phi(\epsilon^{-1}) \equiv \epsilon^{-(1-\hat{\alpha}_\epsilon^{-1})}/\epsilon^{-(1-\alpha^{-1})}, \quad (2.70)$$

and

$$\log[\hat{\phi}(\epsilon^{-1})/\phi(\epsilon^{-1})] = (\alpha^{-1} - \hat{\alpha}_\epsilon^{-1}) \log(\epsilon^{-1}) \rightarrow 0 \quad \text{w.p.1} \quad \text{as} \quad \epsilon \downarrow 0, \quad (2.71)$$

so that

$$\hat{\phi}(\epsilon^{-1})/\phi(\epsilon^{-1}) \rightarrow 1 \quad \text{as} \quad \epsilon \downarrow 0 \quad (2.72)$$

and the FCLT (2.14) holds with the estimator $\hat{\phi}(\epsilon^{-1}) \equiv \epsilon^{-(1-\hat{\alpha}_\epsilon^{-1})}$ used in place of the scaling function $\phi(\epsilon^{-1}) = \epsilon^{-(1-\alpha^{-1})}$. Hence it only remains to estimate the scale parameter Γ . Given that (2.68) holds, the scale parameter is

$$\Gamma = (A/A_\alpha)^{1/\alpha} \quad (2.73)$$

for A in (2.68) and

$$A_\alpha = \left(\int_0^\infty x^{-\alpha} \sin x dx \right)^{-1} = \frac{1-\alpha}{\Gamma(2-\alpha) \cos(\pi\alpha/2)}. \quad (2.74)$$

Hence it suffices to estimate the asymptotic constant A in (2.68). We can estimate in various ways if we estimate the cdf in (2.68) by the empirical cdf.

Hence, under regularity conditions, the sequential stopping rules will again be asymptotically valid. However, in this situation it is often much better to use different (robust) estimators for the mean or to estimate different quantities, such as the median or other percentiles.

Example 4.2.10. (*A counterexample for weak consistency*). Since the SLLN or FWLLN for $\Gamma(t)$ is relatively difficult to establish, it is natural to ask if the weak consistency $\Gamma(t) \Rightarrow \Gamma$ as $t \rightarrow \infty$ in (2.6) might not be enough to ensure asymptotic validity of the sequential stopping rules.

Unfortunately, however, weak consistency of $\Gamma(t)$ is not enough. The difficulty is in establishing the in-probability analog of Theorem 4.2.1 (b).

We now give a direct counterexample. Consider Example 4.2.1 and the process $\Gamma(t)$ defined there. Let N be a unit rate Poisson process independent of $\{X_i : i \geq 1\}$ and let $T_1, T_2, \dots$ be the jump times of the process N . Suppose that

$$\tilde{\Gamma}_{(t)} = \begin{cases} \Gamma(t), & t \notin \cup_{n=1}^{\infty} [T_n, T_n + 1/n), \\ 0, & t \in \cup_{n=1}^{\infty} [T_n, T_n + 1/n). \end{cases} \quad (2.75)$$

Then

$$\begin{aligned} P(\tilde{\Gamma}(t) \neq \Gamma(t)) &= P\left(t \in \left[T_{N(t)}, T_{N(t)} + \frac{1}{N(t)}\right]\right) \\ &\leq P(t - T_{N(t)} \leq \epsilon) + P\left(N(t) \leq \frac{1}{\epsilon}\right) \end{aligned} \quad (2.76)$$

for ϵ arbitrary. Letting $t \rightarrow \infty$, we find that $\limsup_{t \rightarrow \infty} P(\tilde{\Gamma}(t) \leq \Gamma(t)) = 1 - \exp(-\epsilon)$ (recall that the equilibrium age distribution of N is exponential with mean 1). Since ϵ was arbitrary, it follows that $P(\tilde{\Gamma}(t) \neq \Gamma(t)) \rightarrow 0$ as $t \rightarrow \infty$. Then it is evident that $\tilde{\Gamma}(t) \Rightarrow \sigma$ as $t \rightarrow \infty$, since $\Gamma(t) \rightarrow \sigma$ w.p.1 as $t \rightarrow \infty$.

Now, in the setting of Example 4.2.1 using $\tilde{\Gamma}(t)$,

$$\tilde{T}_1(\epsilon) = \inf \left\{ t \geq 0 : z(\delta) \left(\frac{\Gamma(t)}{\sqrt{t}} + a(t) \right) \leq \epsilon \right\}. \quad (2.77)$$

Put $a(t) = 1/t$. Then clearly $z(\delta)(\tilde{\Gamma}(s)/\sqrt{s} + 1/s) \geq z(\delta)/t$ and $s \leq t$, so $\tilde{T}_1(z(\delta)/t) \geq t$. On the other hand, $\tilde{\Gamma}(T_{N(t)+1}) = 0$, so $\tilde{T}_1(z(\delta)/t) \leq T_{N(t)+1}$. By the SLLN, $t^{-1}T_{N(t)+1} \rightarrow 1$ w.p.1 as $t \rightarrow \infty$. Hence $\tilde{T}_1(z(\delta)/t) \sim t$ w.p.1 as $t \rightarrow \infty$. Thus the stopping rule is asymptotically independent of the scaling constant Γ . As a consequence, formation of asymptotically valid confidence intervals is impossible. In fact, even the asymptotic scaling of the rule is incorrect. It is well known that for estimation problems of the type described in Example 4.2.1, the amount of simulation time required to obtain an absolute precision of order ϵ is of order ϵ^{-2} , whereas the stopping rule $\tilde{T}_1(\epsilon)$ based on $\tilde{\Gamma}(t)$ in (2.75) yields a termination time of order ϵ^{-1} .

Chapter 5

Heavy-Traffic Limits for Queues

5.1. Introduction

In this chapter we include additional material on heavy-traffic limits for queues. The first two sections below supplement Chapter 8 in the book; the final section supplements Chapter 9 in the book.

In particular, Section 5.2 discusses general Lévy approximations for queues, obtained by considering a sequence of queueing models, exploiting the FCLT in Section 2.4 above and the continuous-mapping approach. Then Section 5.3 provides the missing proof to Theorem 8.3.1 in the book. Finally, Section 5.4, drawing upon Puhalskii (1994), shows how heavy-traffic limits for arrival, queue-length and departure processes can be used to establish associated limits for waiting-time and workload processes in single-server queues.

5.2. General Lévy Approximations

The Brownian and stable-Lévy approximations for queues in Chapters 5 and 8 in the book are robust approximations: The same approximation, characterized by only a few parameters, serves as an approximation for a large class of queueing models. We obtain the Brownian (stable-Lévy) approximation with light-tailed (heavy-tailed) distributions.

We can obtain a larger, more flexible, class of approximating processes if we consider stochastic-process limits based on a sequence of queueing models, where the input processes are allowed to change with the sequence index. Of course, we also can obtain the previous limit processes in this more general framework, but we can obtain new limit processes as well, which may be useful for applications.

Closely paralleling the previous sections, we can apply the continuous-mapping approach with the reflection map and a Lévy-process FCLT for double sequences in Theorem 2.4.1 here to obtain a stochastic-process limit for workload processes associated with a sequence of queueing models. When we allow the input processes to change in the limit, we can obtain stochastic-process limits without requiring heavy traffic.

As noted in Section 2.4, we obtain a large class of limit processes from the stochastic-process limits for partial sums from double sequences of random variables, with the variables in each sequence being IID. Indeed, the limit process for the net inputs can be an arbitrary Lévy process $\{L(t) : t \geq 0\}$. Of course, in applications it remains to determine the appropriate Lévy process. Since the Lévy process has stationary and independent increments, it suffices to specify the distribution of the random variable $L(1)$, which must be infinitely divisible. From (4.3) in Section 2.4, it suffices to specify the triple (b, σ^2, μ) , where b is the centering constant, σ^2 is the Gaussian coefficient and μ is the Lévy measure. These limiting characteristics can be specified in approximations by exploiting the asymptotic relations in equations (4.10) – (4.12) in Section 2.4.

For applications, it is significant that there is a large class of reflected Lévy processes that are remarkably tractable. In particular, *a reflected Lévy process, constructed from a one-sided reflection, is tractable if the associated Lévy process has no negative jumps*. For example, the steady-state distribution can be characterized by its Laplace transform, which is often called the *generalized Pollaczek-Khintchine transform*, because the Pollaczek-Khintchine transform of the steady-state distribution of the workload process in the M/G/1 queue is a special case.

The original characterization of the steady-state distribution of a reflected Lévy process for the case with no negative jumps is due to Zolotarev (1964); also see Section 24 of Takács (1967), Bingham (1975) and Kella and Whitt (1992b), especially Section 4(a). The short martingale proof in Kella and Whitt (1992b) is convenient.

When a Lévy process L has no negative jumps, the Lévy measure μ concentrates on $(0, \infty)$ and the bilateral Laplace-Stieltjes transform of $L(1)$

is well defined, with *Laplace exponent*

$$\begin{aligned}\psi(s) &\equiv \log Ee^{-sL(1)} \\ &= -bs + \frac{\sigma^2 s^2}{2} + \int_0^\infty (\exp(-sx) - 1 + sh(x))\mu(dx) .\end{aligned}\quad (2.1)$$

An important special case is a subordinator (totally skewed stable Lévy motion with $\beta = 1$ plus a negative drift, which is just (2.1) without the second Brownian term. Storage models with such Lévy net-input processes are analyzed directly in Chapter 4 of Prabhu (1998). With (2.1), we can conveniently characterize the Laplace transform of the steady-state distribution. The following is a generalization of Theorems 5.8.2 and 8.5.2 in the book.

Theorem 5.2.1. (generalized Pollaczek-Khintchine transform) *Let $\{\phi_K(L)(t) : t \geq 0\}$ be a reflected Lévy process, where ϕ_K is the two-sided reflection map, $EL(1) < 0$, L has no negative jumps and L has Laplace exponent ψ in (2.1).*

(a) *If $K = \infty$, then*

$$\lim_{t \rightarrow \infty} P(\phi_K(L))(t) \leq x = H(x) , \quad (2.2)$$

where H is a proper cdf with Laplace-Stieltjes transform

$$\hat{h}(s) \equiv \int_0^\infty e^{-sx} dH(x) = \frac{s\psi'(0)}{\psi(s)} , \quad (2.3)$$

and ψ is the Laplace exponent in (2.1).

(b) *If $K < \infty$, then*

$$\lim_{t \rightarrow \infty} P(\phi_K(L))(t) \leq x = \frac{H(x)}{H(K)} , \quad 0 \leq x \leq K , \quad (2.4)$$

for H in (2.2).

Example 5.2.1. *The special case of the M/G/1 queue.* The workload in unfinished service time in the M/G/1 queue is a reflected Lévy process. If V is a service time and λ is the arrival rate, then the Laplace exponent of the compound-Poisson net-input process is

$$\psi(s) = s - \lambda(1 - E[\exp(-sV)]) .$$

Example 5.2.2. *The gamma process.* A possible subordinator is the gamma process, which can be expressed via the Laplace exponent

$$\psi(s) = \int_0^\infty (e^{-sx} - 1) \frac{e^{-x/\eta}}{x} dx = -\log(1 + \eta s)$$

for constant η ; e.g., see p. 111 of Prabhu (1998). (The centering function is not needed in this case.) If we add a constant negative drift to the gamma process then we obtain a Lévy process with negative drift but without negative jumps, having Laplace exponent $\psi(s) = bs - \log(1 + \eta s)$. If $b > \eta$, then $EL(1) < 0$ and we can apply Theorem 5.2.1. In this case, the steady-state ccdf H^c is easy to compute from its Laplace transform $H^c(s) = [1 - h(s)]/s$ by numerical inversion. The gamma process is a Lévy process without Brownian component; i.e., $b = \sigma^2 = 0$. The Lévy measure has density $\mu(dx) = x^{-1}e^{-x/\eta}$, $x > 0$. We can approximate the gamma process by a compound Poisson process by restricting μ to $[\epsilon, \infty)$ for some $\epsilon > 0$. ■

For other properties of Lévy processes without negative jumps, see Takács (1967), Samorodnitsky and Taqqu (1994), Bertoin (1996) and Prabhu (1998). For a numerical inversion algorithm to calculate first-passage probabilities, see Rogers (2000).

5.3. A Fluid Queue Fed by On-Off Sources

This section is devoted to proving Theorem 8.3.1 in the book, which establishes a FCLT for the cumulative busy time of a single on-off source.

We first restate the theorem. Recall that $B_{n,i}$ is the i^{th} busy period and $I_{n,i}$ is the i^{th} idle period in the n^{th} model, in the sequence of models under consideration. Let

$$\begin{aligned} \mathbf{B}_n(t) &\equiv c_n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} (B_{n,i} - m_{B,n}) \\ \mathbf{I}_n(t) &\equiv c_n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} (I_{n,i} - m_{I,n}) \\ \mathbf{N}_n(t) &\equiv c_n^{-1} [N_n(nt) - \gamma_n nt] \\ \mathbf{B}'_n(t) &\equiv c_n^{-1} [B_n(nt) - \xi_n nt], \quad t \geq 0, \end{aligned} \tag{3.1}$$

where again $\lfloor nt \rfloor$ is the integer part of nt ,

$$\xi_n \equiv \frac{m_{B,n}}{m_{B,n} + m_{I,n}} \quad \text{and} \quad \gamma_n \equiv \frac{1}{m_{B,n} + m_{I,n}} . \quad (3.2)$$

We think of $m_{B,n}$ in (3.1) as the mean busy period, $EB_{n,i}$, and $m_{I,n}$ as the mean idle period, $EI_{n,i}$, in the case $\{(B_{n,i}, I_{n,i}) : i \geq 1\}$ is a stationary sequence for each n , but in general that is not required.

Theorem 5.3.1. (FCLT for the cumulative busy time) *If*

$$(\mathbf{B}_n, \mathbf{I}_n) \Rightarrow (\mathbf{B}, \mathbf{I}) \quad \text{in} \quad (D, M_1)^2 \quad (3.3)$$

for $\mathbf{B}_n$ and $\mathbf{I}_n$ in (3.1), $c_n \rightarrow \infty$, $c_n/n \rightarrow 0$, $m_{B,n} \rightarrow m_B$, $m_{I,n} \rightarrow m_I$, with $0 < m_B + m_I < \infty$, so that $\xi_n \rightarrow \xi$ with $0 \leq \xi \leq 1$ and $\gamma_n \rightarrow \gamma > 0$ for ξ_n and γ_n in (3.2), and

$$P(\text{Disc}(\mathbf{B}) \cap \text{Disc}(\mathbf{I}) = \phi) = 1 , \quad (3.4)$$

then

$$(\mathbf{B}_n, \mathbf{I}_n, \mathbf{N}_n, \mathbf{B}'_n) \Rightarrow (\mathbf{B}, \mathbf{I}, \mathbf{N}, \mathbf{B}') \quad \text{in} \quad (D, M_1)^4 , \quad (3.5)$$

for $\mathbf{N}_n, \mathbf{B}'_n$ in (3.1) and

$$\begin{aligned} \mathbf{N}(t) &\equiv -\gamma[\mathbf{B}(\gamma t) + \mathbf{I}(\gamma t)] \\ \mathbf{B}'(t) &\equiv (1 - \xi)\mathbf{B}(\gamma t) - \xi\mathbf{I}(\gamma t) . \end{aligned} \quad (3.6)$$

The possibility of the limit processes having discontinuous sample paths makes the required argument more complicated than what it might otherwise be. To make that clear, before presenting an argument that works, we present two false starts.

5.3.1. Two False Starts

For the first false start, note that the cumulative busy-time process can be bounded above and below by random sums by

$$c_n^{-1} \sum_{i=1}^{N_n(nt)} B_{n,i} \leq c_n^{-1} B_n(nt) \leq c_n^{-1} \sum_{i=1}^{N_n(nt)+1} B_{n,i} , \quad (3.7)$$

so let us start by trying to find limits for the outer terms in (3.7). We apply the continuous mapping theorem with addition (Section 12.7 in the

book) and the inverse map (Sections 13.7 and 13.8 in the book) to get, first, $\mathbf{B}_n + \mathbf{I}_n \Rightarrow \mathbf{B} + \mathbf{I}$ and then $\mathbf{N}_n \Rightarrow \mathbf{N}$ jointly.

As a consequence, we get $\mathbf{T}_n \Rightarrow \gamma e$, where

$$T_n(t) \equiv n^{-1}N_n(nt), \quad t \geq 0. \quad (3.8)$$

Then we try to treat the term on the left in (3.7) by writing

$$\begin{aligned} & c_n^{-1} \left(\sum_{i=1}^{N_n(nt)} B_{n,i} - m_{n,2} \gamma_n nt \right) \\ &= c_n^{-1} \left(\sum_{i=1}^{\lfloor nt \rfloor} B_{n,i} - m_{n,2} \right) \circ \frac{N_n(nt)}{n} + m_{n,2} (c_n^{-1} [N_n(nt) - \gamma_n nt]) \\ &\Rightarrow \mathbf{B}(\gamma t) - m_2(\gamma[\mathbf{B}(\gamma t) + \mathbf{I}(\gamma t)]) = (1 - \xi)\mathbf{B}(\gamma t) - \xi\mathbf{I}(\gamma t). \end{aligned} \quad (3.9)$$

This argument works fine if $P(\mathbf{B} \in C) = 1$, but not otherwise. This argument is not valid here because we need to apply addition when the limit processes $\mathbf{B} \circ \gamma e$ and $-\gamma(\mathbf{B} \circ \gamma e + \mathbf{I} \circ \gamma e)$ typically have common discontinuities of opposite sign. (If they had the same sign, then we could apply Theorem 12.7.3 in the book.) Hence we need to find a different approach.

For our second false start, instead of (3.7), we find different bounds for the cumulative busy-time process, in particular, note that

$$\begin{aligned} \mathbf{B}'_n(t) &\leq c_n^{-1} \left[(1 - \xi_n) \sum_{i=1}^{N_n(nt)+1} B_{n,i} - \xi_n \sum_{i=1}^{N_n(nt)} I_{n,i} \right] \\ &\leq c_n^{-1} \left[(1 - \xi_n) \sum_{i=1}^{N_n(nt)+1} (B_{n,i} - m_{n,1}) - \xi_n \sum_{i=1}^{N_n(nt)} (I_{n,i} - m_{n,2}) \right] \\ &\quad + c_n^{-1} m_{n,1} \\ \mathbf{B}'_n(t) &\geq c_n^{-1} \left[(1 - \xi_n) \sum_{i=1}^{N_n(nt)} B_{n,i} - \xi_n \sum_{i=1}^{N_n(nt)+1} I_{n,i} \right] \\ &\geq c_n^{-1} \left[(1 - \xi_n) \sum_{i=1}^{N_n(nt)} (B_{n,i} - m_{n,1}) - \xi_n \sum_{i=1}^{N_n(nt)+1} (I_{n,i} - m_{n,2}) \right] \\ &\quad - c_n^{-1} m_{n,2}. \end{aligned}$$

Note that the deterministic terms $c_n^{-1}m_{n,1}$ and $c_n^{-1}m_{n,2}$ are asymptotically negligible. Thus, let the asymptotically bounding processes be

$$\mathbf{B}_n^u(t) \equiv c_n^{-1} \left[(1 - \xi_n) \sum_{i=1}^{N_n(nt)+1} (B_{n,i} - m_{n,1}) - \xi_n \sum_{i=1}^{N_n(nt)} (I_{n,i} - m_{n,2}) \right] \quad (3.10)$$

and

$$\mathbf{B}_n^l(t) \equiv c_n^{-1} \left[(1 - \xi_n) \sum_{i=1}^{N_n(nt)} (B_{n,i} - m_{n,1}) - \xi_n \sum_{i=1}^{N_n(nt)+1} (I_{n,i} - m_{n,2}) \right]. \quad (3.11)$$

Also let

$$\mathbf{T}_n(t) \equiv \frac{N_n(nt)}{n}, \quad \mathbf{T}'_n(t) \equiv \frac{N_n(nt) + 1}{t} \quad (3.12)$$

and

$$\mathbf{N}'_n(t) \equiv c_n^{-1} [N_n(nt) + 1 - \gamma_n nt], \quad t \geq 0. \quad (3.13)$$

As before, we apply the continuous mapping theorem with the addition and the inverse map to get, first $\mathbf{B}_n + \mathbf{I}_n \Rightarrow \mathbf{B} + \mathbf{I}$ and then $\mathbf{N}_n \Rightarrow \mathbf{N}$ and $\mathbf{N}'_n \Rightarrow \mathbf{N}$, all jointly. Given $\mathbf{N}_n \Rightarrow \mathbf{N}$ and $\mathbf{N}'_n \Rightarrow \mathbf{N}$ we obtain $\mathbf{T}_n \Rightarrow \gamma e$ and $\mathbf{T}'_n \Rightarrow \gamma e$ by multiplying by c_n/n . Applying the composition map, we obtain

$$\mathbf{B}_n^u = (1 - \xi_n)\mathbf{B}_n \circ \mathbf{T}'_n - \xi_n \mathbf{I}_n \circ \mathbf{T}_n \Rightarrow \mathbf{B}' \quad (3.14)$$

and

$$\mathbf{B}_n^l = (1 - \xi_n)\mathbf{B}_n \circ \mathbf{T}_n - \xi_n \mathbf{I}_n \circ \mathbf{T}'_n \Rightarrow \mathbf{B}', \quad (3.15)$$

again jointly with the other limits. Hence we are close to obtaining (3.5). However, even though $(\mathbf{B}_n^l, \mathbf{B}_n^u) \Rightarrow (\mathbf{B}', \mathbf{B}')$ and $\mathbf{B}_n^l \leq \mathbf{B}'_n \leq \mathbf{B}_n^u$, we cannot deduce that $\mathbf{B}'_n \Rightarrow \mathbf{B}'$ in (D, M_1) .

5.3.2. The Proof

We can deduce that $\mathbf{B}'_n \Rightarrow \mathbf{B}'$ in the weaker Skorohod M_2 topology by this reasoning, though, by virtue of Corollary 12.11.4 in the book, from which we can deduce convergence of the finite-dimensional distributions. To get the desired M_1 limit, it thus suffices to apply Theorem 12.5.1 (iv) in the book and control the oscillations as in equation (12.5.3) of the book. To do

so, we introduce a slightly different approximation. Let

$$\mathbf{B}_n^a(t) = c_n^{-1} \left[(1 - \xi_n) \sum_{i=1}^{N_n^B(nt)} (B_{n,i} - m_{n,1}) - \xi_n \sum_{i=1}^{N_n^I(nt)} (I_{n,i} - m_{n,2}) \right], \quad (3.16)$$

where $N_n^B(t)$ and $N_n^I(t) = N_n(t)$ are the number of complete busy periods and idle periods by time t . Reasoning as with (3.14) and (3.15) we can deduce that $\mathbf{B}_n^a \Rightarrow \mathbf{B}'$. However, we can make a stronger connection between $\mathbf{B}_n^a$ and $\mathbf{B}_n$. Note that

$$N_n^I(t) = N_n(t) \leq N_n^B(t) \leq N_n(t) + 1$$

and

$$\mathbf{B}_n^a \circ \mathbf{S}_n = \mathbf{B}_n \circ \mathbf{S}_n \quad \text{and} \quad \mathbf{B}_n^a \circ \mathbf{S}'_n = \mathbf{B}_n \circ \mathbf{S}'_n$$

where

$$\mathbf{S}_n(t) \equiv n^{-1} \tau_{n, \lfloor nt \rfloor}, \quad \mathbf{S}'_n(t) \equiv n^{-1} \tau'_{n, \lfloor nt \rfloor},$$

$$\tau_{n,0} = 0,$$

$$\tau_{n,k} \equiv B_{n,1} + I_{n,1} + \cdots + B_{n,k} + I_{n,k}, \quad k \geq 1,$$

and

$$\tau'_{n,k} \equiv \tau_{n,k} + B_{n,k+1}, \quad k \geq 0.$$

Moreover $\mathbf{B}_n^a$ is piecewise-constant and $\mathbf{B}_n$ is piecewise linear in each of the intervals $[n^{-1} \tau_{n,k}, n^{-1} \tau'_{n,k}]$ and $[n^{-1} \tau'_{n,k}, n^{-1} \tau_{n,k+1}]$. Hence we can relate the oscillation of $\mathbf{B}_n$ to those of $\mathbf{B}_n^a$.

First, we can apply the Skorohod representation theorem to replace convergence in distribution by convergence w.p.1. We obtain $\mathbf{B}_n^a \rightarrow \mathbf{B}'$ w.p.1 for new versions of these processes. From the specific structure above, we can construct the corresponding special version of $\mathbf{B}'_n$ associated with $\mathbf{B}_n^a$. (It is the piecewise-linear interpolation of the piecewise-constant function.) Since $\mathbf{B}_n^a \rightarrow \mathbf{B}'$, $\mathbf{S}_n \rightarrow \gamma^{-1}e$ and $\mathbf{S}'_n \rightarrow \gamma^{-1}e$ for the new versions, we can deduce that $\mathbf{B}'_n(t) \rightarrow \mathbf{B}'(t)$ w.p.1 for each continuity point t of $\mathbf{B}'$. (We also got this part from the convergence of $\mathbf{B}_n^l$ and $\mathbf{B}_n^u$.) Let w_s be the M_1 oscillation function over the interval $[0, T]$, where T is chosen to be a continuity point of $\mathbf{B}'$, i.e.,

$$w_s(x, \delta) \equiv \sup_{0 \vee (t-\delta) \leq t_1 < t_2 < t_3 \leq (t+\delta) \wedge T} \{|x(t_2) - [x(t_1), x(t_3)]|\}$$

where $[x(t_1), x(t_3)]$ is the line segment connecting $x(t_1)$ and $x(t_3)$. From the properties above, we can deduce that

$$w_s(\mathbf{B}'_n, \delta) \leq w_s(\mathbf{B}_n^a, 2\delta)$$

for all suitably large n . Since $\mathbf{B}_n^a \rightarrow \mathbf{B}'$, we deduce that

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w_s(\mathbf{B}_n^a, \delta) = 0, \quad (3.17)$$

which implies the same limit with $\mathbf{B}_n^a$ replaced by $\mathbf{B}'_n$ in (3.17). By the characterization of M_1 convergence in Theorem 12.5.1 (iv) in the book, we get $\mathbf{B}'_n \rightarrow \mathbf{B}'$ w.p.1 (in D, M_1) for the special versions and thus $\mathbf{B}'_n \Rightarrow \mathbf{B}'$ for the original versions. This can be done jointly with the other processes, so that we get (3.5).

5.4. From Queue Lengths to Waiting Times

In this section, following Puhalskii (1994), we show how the continuous-mapping approach with the inverse map and nonlinear centering term, Theorem 13.7.4 in the book, can be used to convert limits for arrival, departure and queue-length processes into associated limits for waiting-time and workload processes in quite general queueing models. The nonlinear centering enables us to capture nonstationary phenomena.

5.4.1. The Setting

The setting is a family of queueing models indexed by n . Suppose that all arrivals eventually get served and then depart, so that the queue length (number of customers in the system) at time t is just the initial queue length plus the arrivals minus the departures, i.e.,

$$Q_n(t) = Q_n(0) + A_n(t) - D_n(t), \quad t \geq 0, \quad (4.1)$$

where $Q_n(t)$ is the queue length at time t , $A_n(t)$ is the number of arrivals in the interval $[0, t]$, and $D_n(t)$ is the number of departures in the interval $[0, t]$, all in model n . To treat customer waiting times (but not the workload), we need to make assumptions about the service mechanism. In particular, we assume that the customers are served one at a time in order of their arrival. Thus, we are again in the setting of the standard single-server queue. Let $A_n(t)$ count the new arrivals, and let $D_n(t)$ counts all departures, including those customers originally in the system at time 0. Note that $\{A_n(t) : t \geq 0\}$ and $\{D_n(t) : t \geq 0\}$ are counting processes. As a regularity condition, we assume that $A_n(0) = D_n(0) = 0$.

5.4.2. The Inverse Map with Nonlinear Centering

We can use the inverse map to define related quantities of interest. Let $A_{n,k}$ be the arrival time of the k^{th} arriving customer, $D_{n,k}$ the departure time of the k^{th} arriving customer and $L_n(t)$ the workload facing the server at time t , not counting arrivals after time t (the virtual waiting time), all in model n . Then

$$\begin{aligned} A_{n,k} &\equiv \inf\{s \geq 0 : A_n(s) > (k-1)^+\} , \\ D_{n,k} &\equiv \inf\{s \geq 0 : D_n(s) > (Q_n(0) + k - 1)^+\} , \\ L_n(t) &\equiv \inf\{s \geq 0 : D_n(s) > Q_n(0) + A_n(t)\} \end{aligned} \quad (4.2)$$

for $k \geq 1$ and $t \geq 0$, where $(x)^+ = \max\{x, 0\}$.

Let $W_{n,k}$ be the waiting time for arriving customer k to begin service and let $W'_{n,k}$ be the waiting time until customer k completes service. Then, under the assumptions about the service mechanism above,

$$W_{n,k} \equiv [D_{n,k-1} - A_{n,k}]^+ , \quad (4.3)$$

and

$$W'_{n,k} \equiv D_{n,k} - A_{n,k}, \quad k \geq 1 . \quad (4.4)$$

Suppose that the time scaling is already incorporated in the models indexed by n . We assume that functional weak laws of large numbers (FWLLNs) holds with additional space scaling by n and that FCLTs hold with additional space scaling by c_n after centering. Thus, let

$$\begin{aligned} \hat{\mathbf{X}}_n(t) &\equiv n^{-1}D_n(t), \\ \hat{\mathbf{Y}}_n(t) &\equiv n^{-1}A_n(t), \\ \hat{\mathbf{Q}}_n(t) &\equiv n^{-1}Q_n(t), \\ \mathbf{X}_n(t) &\equiv c_n(\hat{\mathbf{X}}_n - \mathbf{x}), \\ \mathbf{Y}_n(t) &\equiv c_n(\hat{\mathbf{Y}}_n - \mathbf{y}), \\ \mathbf{Q}_n(t) &\equiv c_n(\hat{\mathbf{Q}}_n - \mathbf{q}), \quad t \geq 0. \end{aligned} \quad (4.5)$$

We assume that

$$(\hat{\mathbf{X}}_n, \hat{\mathbf{Y}}_n, \hat{\mathbf{Q}}_n) \Rightarrow (\mathbf{x}, \mathbf{y}, \mathbf{q}) \quad \text{in} \quad (D^3, WM_1) \quad (4.6)$$

where $\mathbf{x}, \mathbf{y} \in D_{\uparrow}$, $\mathbf{q} \in D$ and, by (4.1),

$$\mathbf{q}(t) = \mathbf{q}(0) + \mathbf{y}(t) - \mathbf{x}(t), \quad t \geq 0 . \quad (4.7)$$

We will also impose smoothness conditions on $\mathbf{x}$ and $\mathbf{y}$. In addition, we assume that $c_n \rightarrow \infty$ and

$$(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Q}_n) \Rightarrow (\mathbf{X}, \mathbf{Y}, \mathbf{Q}) \quad \text{in} \quad (D^3, WM_1). \quad (4.8)$$

As a consequence of (4.1) and (4.5)–(4.8),

$$\mathbf{Q}(t) - \mathbf{Q}(0) = \mathbf{A}(t) - \mathbf{D}(t) \quad \text{for} \quad t > 0. \quad (4.9)$$

Given the FWLLN (4.6) and the FCLT (4.8), we want to establish related limits for appropriately scaled versions of the random variables $A_{n,k}$, $D_{n,k}$, $L_n(t)$, $W_{n,k}$ and $W'_{n,k}$ in (4.2)–(4.4). For that purpose, let

$$\hat{\mathbf{D}}_n(t) \equiv D_{n, \lfloor nt \rfloor}, \quad \hat{\mathbf{A}}_n(t) \equiv A_{n, \lfloor nt \rfloor}, \quad \hat{\mathbf{L}}_n(t) = L_n(t) \quad (4.10)$$

and

$$\hat{\mathbf{W}}_n(t) \equiv W_{n, \lfloor nt \rfloor} \quad \text{and} \quad \hat{\mathbf{W}}'_n(t) \equiv W'_{n, \lfloor nt \rfloor}, \quad t \geq 0. \quad (4.11)$$

We now form the final scaled random elements of D . Let

$$\begin{aligned} \mathbf{U}_n(t) &\equiv c_n(\hat{\mathbf{X}}_n^{-1} - \mathbf{x}^{-1}), \\ \mathbf{V}_n(t) &\equiv c_n(\hat{\mathbf{Y}}_n^{-1} - \mathbf{y}^{-1}), \\ \mathbf{A}_n(t) &\equiv c_n(\hat{\mathbf{A}}_n - \mathbf{y}^{-1}), \\ \mathbf{D}_n(t) &\equiv c_n(\hat{\mathbf{D}}_n - \mathbf{x}^{-1} \circ \mathbf{z}_1), \\ \mathbf{L}_n(t) &\equiv c_n(\hat{\mathbf{L}}_n - \mathbf{x}^{-1} \circ \mathbf{z}_2), \\ \mathbf{W}_n(t) &\equiv c_n(\hat{\mathbf{W}}_n - (\mathbf{x}^{-1} \circ \mathbf{z}_1 - \mathbf{y}^{-1})), \\ \mathbf{W}'_n(t) &\equiv c_n(\hat{\mathbf{W}}'_n - (\mathbf{x}^{-1} \circ \mathbf{z}_1 - \mathbf{y}^{-1})), \quad t \geq 0. \end{aligned} \quad (4.12)$$

We now state the theorem.

Theorem 5.4.1. (FCLT for the workload and waiting time given a FCLT for arrivals, departures and queue length) *Suppose that the limit (4.8) holds for $\mathbf{X}_n$, $\mathbf{Y}_n$, $\mathbf{Q}_n$ in (4.5), where $c_n \rightarrow \infty$, $\mathbf{x}, \mathbf{y} \in \Lambda$ and are absolutely continuous with continuous positive derivatives $\dot{\mathbf{x}}$, $\dot{\mathbf{y}}$, and $P(\mathbf{X}(0) = 0) = P(\mathbf{Y}(0) = 0) = 1$. Then, jointly with (4.8),*

$$(\mathbf{U}_n, \mathbf{V}_n, \mathbf{A}_n, \mathbf{D}_n) \Rightarrow (\mathbf{U}, \mathbf{V}, \mathbf{A}, \mathbf{D}) \quad (4.13)$$

in (D^4, WM_1) for $\mathbf{U}_n$, $\mathbf{V}_n$, $\mathbf{A}_n$ and $\mathbf{D}_n$ in (4.12), where

$$\mathbf{U} = \frac{-\mathbf{X} \circ \mathbf{x}^{-1}}{\dot{\mathbf{x}} \circ \mathbf{x}^{-1}}, \quad \mathbf{V} = \mathbf{A} = \frac{-\mathbf{Y} \circ \mathbf{y}^{-1}}{\dot{\mathbf{y}} \circ \mathbf{y}^{-1}} \quad (4.14)$$

and

$$\mathbf{D} = \frac{-\mathbf{X} \circ \mathbf{x}^{-1} \circ \mathbf{z}_1 + \mathbf{Q}(0)\mathbf{1}}{\dot{\mathbf{x}} \circ \mathbf{x}^{-1} \circ \mathbf{z}_1}, \quad \mathbf{z}_1 = \mathbf{q}(0)\mathbf{1} + \mathbf{e}, \quad (4.15)$$

where $\mathbf{e}(t) = t$ for $t \geq 0$. If, in addition,

$$P(\text{Disc}(\mathbf{X} \circ \mathbf{x}^{-1} \circ \mathbf{z}_2) \cap \text{Disc}(\mathbf{Y}) = \phi) = 1 \quad (4.16)$$

for

$$\mathbf{z}_2 = \mathbf{q}(0)\mathbf{1} + \mathbf{y}, \quad (4.17)$$

then, jointly with (4.8) and (4.13),

$$\mathbf{L}_n \Rightarrow \mathbf{L} \quad (4.18)$$

for $\mathbf{L}_n$ in (4.12), where

$$\mathbf{L} = \frac{-\mathbf{X} \circ \mathbf{x}^{-1} \circ \mathbf{z}_2 + \mathbf{Y} + \mathbf{Q}(0)\mathbf{1}}{\dot{\mathbf{x}} \circ \mathbf{x}^{-1} \circ \mathbf{z}_2}. \quad (4.19)$$

If, in addition,

$$P(\text{Disc}(\mathbf{A}) \cap \text{Disc}(\mathbf{D}) = \phi) = 1, \quad (4.20)$$

then, jointly with (4.8), (4.13) and (4.18),

$$(\mathbf{W}_n, \mathbf{W}'_n) \Rightarrow (\mathbf{D} - \mathbf{A}, \mathbf{D} - \mathbf{A}) \quad (4.21)$$

in (D^2, WM_1) for $\mathbf{W}_n$ and $\mathbf{W}'_n$ in (4.12).

In preparation for the proof, we now restate Theorem 13.7.4 from the book. Recall that $D_\uparrow$ is the subset of all nondecreasing nonnegative functions in D . Recall that D_u is the subset of all functions in $D([0, \infty), \mathbb{R})$ that are unbounded above and satisfy $x(0) \geq 0$.

The following is Puhalskii's (1994) result extended to allow discontinuous limits.

Theorem 5.4.2. Suppose that $x_n \in D_u$, $y_n \in D_\uparrow$, $c_n \rightarrow \infty$,

$$c_n(x_n - x, y_n - y) \rightarrow (u, v) \quad \text{in } D \times D \quad (4.22)$$

with one of the J_1 , M_1 or M_2 topologies, where $u(0) = 0$, u has no positive jumps if the topology is J_1 ,

$$\text{Disc}(u \circ x^{-1} \circ y) \cap \text{Disc}(v) = \phi, \quad (4.23)$$

$y \in C_{\uparrow\uparrow}$ and x is absolutely continuous with a continuous positive derivative $\dot{x}$, then

$$c_n(x_n^{-1} \circ y_n - x^{-1} \circ y) \rightarrow \frac{v - u \circ x^{-1} \circ y}{\dot{x} \circ x^{-1} \circ y} \quad \text{in } D \quad (4.24)$$

with the same topology.

Proof of Theorem 5.4.1. We start by applying the Skorohod representation theorem to replace convergence in distribution by convergence w.p.1. For simplicity, we do not introduce new notation for these special versions of the random functions converging w.p.1. Thus consider a single sample path for which the limit (4.8) holds. Now we can apply the deterministic convergence-preservation results. From (4.2)–(4.11), we see that we can represent $\hat{\mathbf{D}}_n$, $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{L}}_n$ in terms of $\hat{\mathbf{X}}_n$ and $\hat{\mathbf{Y}}_n$ via the inverse map

$$\begin{aligned}\hat{\mathbf{A}}_n(t) &\equiv \inf\{s \geq 0 : A_n(s) > \lfloor nt \rfloor - 1\} \\ &= \inf\{s \geq 0 : \hat{\mathbf{Y}}_n(s) > (\lfloor nt \rfloor - 1)/n\} \\ &= (\hat{\mathbf{Y}}_n^{-1} \circ \xi_n)(t), \quad t \geq 0,\end{aligned}\tag{4.25}$$

where

$$\xi_n(t) = (\lfloor nt \rfloor - 1)^+/n, \quad t \geq 0,\tag{4.26}$$

$$\begin{aligned}\hat{\mathbf{D}}_n(t) &\equiv \inf\{s \geq 0 : D_n(s) > (Q_n(0) + \lfloor nt \rfloor - 1)^+\} \\ &= \inf\{s \geq 0 : \hat{\mathbf{X}}_n(s) > \{Q_n(0) + \lfloor nt \rfloor - 1\}^+/n\} \\ &= (\hat{\mathbf{X}}_n^{-1} \circ \zeta_n)(t), \quad t \geq 0,\end{aligned}\tag{4.27}$$

where

$$\zeta_n(t) = (Q_n(0) + \lfloor nt \rfloor - 1)^+/n, \quad t \geq 0,\tag{4.28}$$

and

$$\begin{aligned}\hat{\mathbf{L}}_n(t) &\equiv \inf\{s \geq 0 : D_n(s) > Q_n(0) + A_n(nt)\} \\ &= \inf\{s \geq 0 : \hat{\mathbf{X}}_n(s) > \hat{\mathbf{Q}}_n(0) + \hat{\mathbf{Y}}_n(t)\} \\ &= [\hat{\mathbf{X}}_n^{-1} \circ (\hat{\mathbf{Q}}_n(0)\mathbf{1} + \hat{\mathbf{Y}}_n)](t), \quad t \geq 0\end{aligned}\tag{4.29}$$

where $\mathbf{1}(t) \equiv 1$, $t \geq 0$. Given (4.3)–(4.27),

$$\hat{\mathbf{W}}_n(t) = [(\hat{\mathbf{D}}_n \circ \xi_n)(t) - \hat{\mathbf{A}}_n(t)]^+, \quad t \geq 0,\tag{4.30}$$

for ξ_n in (4.26) and

$$\hat{\mathbf{W}}_n'(t) = (\hat{\mathbf{D}}_n - \hat{\mathbf{A}}_n)(t), \quad t \geq 0.\tag{4.31}$$

We now return to the proof of (4.13). First, for the inverse processes $\hat{\mathbf{X}}_n^{-1}$ and $\hat{\mathbf{Y}}_n^{-1}$, we apply Theorem 13.7.2 from the book. Given those two limits, we treat $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{D}}_n$ by applying the composition result, Theorem 12.3.1. Alternatively, we directly apply Theorem 5.4.2 above, noting that $\xi_n \rightarrow \mathbf{e}$,

$\zeta_n \rightarrow \mathbf{z}_1$, $c_n(\xi_n - e) \rightarrow \mathbf{0}$ and $c_n(\zeta_n - z_1) \Rightarrow \hat{\mathbf{Q}}(0)\mathbf{1}$. To treat $\hat{\mathbf{L}}_n$ we again apply Theorem 12.3.1 or Theorem 5.4.2 above, using the fact that $\hat{\mathbf{Q}}_n(0)\mathbf{1} + \hat{\mathbf{Y}}_n \rightarrow \mathbf{z}_2$ and $c_n(\hat{\mathbf{Q}}_n(0)\mathbf{1} + \hat{\mathbf{Y}}_n - \mathbf{z}_2) \rightarrow \mathbf{Y} + \mathbf{Q}(0)\mathbf{1}$ in D . Finally, to treat $\hat{\mathbf{W}}_n$ and $\hat{\mathbf{W}}'_n$, we use the subtraction map. We first apply subtraction directly to $\hat{\mathbf{W}}'_n$ in (4.31). Since $\xi_n \Rightarrow \mathbf{e}$, we can conclude that $\mathbf{W}_n$ has the same limit as $\mathbf{W}'_n$. ■

Remark 5.4.1. If Theorem 5.4.1 holds for stationary models, then $\mathbf{x} = \mathbf{y} = \lambda\mathbf{e}$, and $\mathbf{q} = \mathbf{q}(0)\mathbf{1}$. Suppose in addition that $\mathbf{q}(0) = \mathbf{0}$. By (4.9), if we cannot conclude that the limit processes almost surely have continuous paths, then we should anticipate $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Q}$ can have common discontinuities. Then

$$\mathbf{U} = -\lambda^{-1}\mathbf{X} \circ \lambda^{-1}\mathbf{e} \quad (4.32)$$

and

$$\mathbf{V} = \mathbf{A} = -\lambda^{-1}\mathbf{Y} \circ \lambda^{-1}\mathbf{e}. \quad (4.33)$$

Condition (4.16) then becomes

$$P(\text{Disc}(\mathbf{X}) \cap \text{Disc}(\mathbf{Y}) = \emptyset) = 1 \quad (4.34)$$

and

$$\mathbf{D} = \lambda^{-1}(\mathbf{Y} - \mathbf{X} + \mathbf{Q}(0)\mathbf{1}) = \lambda^{-1}\mathbf{Q}. \quad (4.35)$$

Then the centering terms in (4.21) become

$$\mathbf{x}^{-1} \circ \mathbf{z}_1 - \mathbf{y}^{-1} = \lambda^{-1}\mathbf{e} - \lambda^{-1}\mathbf{e} = \mathbf{0} \quad (4.36)$$

and

$$\begin{aligned} \mathbf{D} - \mathbf{A} &= \lambda^{-1}(\mathbf{Y} - \mathbf{X} + \mathbf{Q}(0)\mathbf{1}) + \lambda^{-1} \circ \mathbf{X} \circ \lambda^{-1}\mathbf{e} \\ &= \lambda^{-1}(\mathbf{Q} + \mathbf{X} \circ \lambda^{-1}\mathbf{e}). \end{aligned}$$

■

5.4.3. An Application to Central-Server Models

Following Puhalskii (1994), we illustrate how Theorem 5.4.1 can be applied by considering a limit for a central-server model. Central-server models were originally introduced to model the contention among programs for the processor and input-output devices in a multiprogrammed computer system; e.g., see Section 3.4.2 of Lavenberg and Sauer (1983). The specific model we consider is a closed queueing network with $n + 1$ single-server queues,

one of which is called the central-server queue while the others are called peripheral queues. There are n customers (jobs) in the network, one for each peripheral queue. Each customer has a designated distinct peripheral queue. Each customer circulates between the central-server queue and its own designated peripheral queue. The customers are served one at a time in order of arrival at the central-server queue. The service times are assumed to be mutually independent exponential random variables. (That ensures that the closed network has a product-form steady-state distribution.) Let the mean service time at each peripheral queue be λ^{-1} , and let the mean service time at the central-server queue be $(n\mu)^{-1}$.

Since only one customer receives service at each peripheral queue, there is no contention there. Thus, each customer enters service at its peripheral immediately upon arrival. Consequently, the $(n+1)$ -queue model is equivalent to a 2-queue model, with one queue being the central-server queue and the other queue being an infinite-server queue. Moreover, the number of customers at the central-server queue evolves as a birth-and-death process with state-dependent transition rates. Let $Q_n(t)$ denote the number of customers at the central-server queue at time t , as a function of n . When $Q_n(t) = k$, the birth (arrival) rate is $(n-k)\lambda$ and the death (service) rate is $n\mu$. Hence the steady-state distribution is easy to calculate.

However, it is also of interest to consider limits as $n \rightarrow \infty$ in order to better understand the behavior of such systems with fast central servers and many customers. First a FLLN is quite elementary. For that purpose, let $A_n(t)$ and $D_n(t)$ count the numbers of arrivals and departures, respectively, at the central-server queue in the interval $[0, t]$. Then form the scaled processes $\hat{\mathbf{X}}_n$, $\hat{\mathbf{Y}}_n$ and $\hat{\mathbf{Q}}_n$ as in (4.5). It is then relatively elementary to show that, if $\hat{\mathbf{Q}}_n(0) = \mathbf{q}(0)$, $0 \leq q(0) \leq 1$, then the FWLLN in (4.6) holds here with

$$\mathbf{x}(t) = \mu t, \quad \mathbf{y}(t) = \lambda \int_0^t [1 - \mathbf{q}(s)] ds \quad (4.37)$$

and $\mathbf{q}$ satisfying the ordinary differential equation

$$\dot{\mathbf{q}}(t) \equiv \frac{d\mathbf{q}}{dt}(t) = \lambda(1 - \mathbf{q}(t)) - \mu. \quad (4.38)$$

Kogan, Lipster and Smorodinskii (1986) then established the following result; also see Chapter 8, Section 3, of Liptser and Shiryaev (1989) and Puhalskii (1994).

Theorem 5.4.3. (FCLT for the central-server model) *If*

$$\sqrt{n}[\mathbf{Q}_n(0) - \mathbf{q}(0)] \Rightarrow \mathbf{Q}(0) \quad \text{in } \mathbb{R}, \quad (4.39)$$

then the joint limit (4.8) holds with

$$\mathbf{X}(t) = \sqrt{\mu}\mathbf{B}_2(t) , \quad (4.40)$$

$$\mathbf{Y}(t) = \int_0^t \sqrt{\lambda(1 - \mathbf{q}(s))} d\mathbf{B}_1(s) - \lambda \int_0^t \mathbf{Q}(s) ds \quad (4.41)$$

and

$$\mathbf{Q}(t) = \mathbf{Q}(0) + \mathbf{X}(t) - \mathbf{Y}(t), \quad t \geq 0 . \quad (4.42)$$

The limit process $\mathbf{Q}$ can be expressed as the solution to

$$\begin{aligned} \mathbf{Q}(t) = \mathbf{Q}(0) & - \lambda \int_0^t \mathbf{Q}(s) ds \\ & + \int_0^t \sqrt{\lambda(1 - \mathbf{q}(s))} d\mathbf{B}_1(s) - \sqrt{\mu}\mathbf{B}_2(t) . \end{aligned} \quad (4.43)$$

We can now combine Theorems 5.4.1 and 5.4.3 to obtain associated limits for the scaled versions of $\hat{\mathbf{A}}_n, \hat{\mathbf{D}}_n, \hat{\mathbf{L}}_n$ in (4.10) and $\hat{\mathbf{W}}_n$ and $\hat{\mathbf{W}}_n'$ in (4.11), as stated in Theorem 5.4.1. Theorem 5.4.1 is genuinely helpful here, because these limits are not so easy to obtain directly.

Theorem 5.4.1 has also been applied by Mandelbaum, Massey, Reiman and Stolyar (1999).

Chapter 6

The Space D

6.1. Introduction

This chapter contains proofs omitted from Chapter 12 of the book, with the same title. For convenience, the theorems are restated here. The section and theorem numbers parallel Chapter 12 of the book, so the proofs should be easy to find.

Here is how the present chapter is organized: We start in Section 6.2 by discussing regularity properties of the function space D . A key property, which we frequently use, is the fact that any function in D can be approximated uniformly closely by piecewise-constant functions with only finitely many discontinuities.

In Section 6.3 we introduce the strong and weak versions of the M_1 topology on $D([0, T], \mathbb{R}^k)$, referred to as SM_1 and WM_1 , and establish basic properties. We also discuss the relation among the non-uniform Skorohod topologies on D . In Section 6.4 we discuss local uniform convergence at continuity points and relate it to oscillation functions used to characterize different forms of convergence.

In Section 6.5 we provide several different alternative characterizations of SM_1 and WM_1 convergence. Some involve parametric representations of the completed graphs and others involve oscillation functions. It is significant that there are forms of the oscillation-function characterizations that involve considering one function argument t at a time. Consequently, the examples in Figure 11.2 of the book tend to be more than illustrative: The topologies are characterized by the local behavior in the neighborhood of single discontinuities.

In Section 6.6 we discuss conditions that allow us to strengthen the mode of convergence from WM_1 to SM_1 . The key condition is to have the

coordinate limit functions have no common discontinuities. In Section 6.7 we study how SM_1 convergence in $D([0, T], \mathbb{R}^k)$ can be characterized by associated limits of mappings.

In Section 6.8 we exhibit a complete metric topologically equivalent to the incomplete metric inducing the SM_1 topology introduced earlier. As with the J_1 metric d_{J_1} in equation (3.2) of Section 3.3 in the book, the natural M_1 metric is incomplete, but there exists a topologically equivalent complete metric, so that D with the SM_1 topology is Polish (metrizable as a complete separable metric space).

In Section 6.9 we discuss extensions of the SM_1 and WM_1 topologies on $D([0, T], \mathbb{R}^K)$ to corresponding spaces of functions with non-compact domains. The principal example of such a non-compact domain is the interval $[0, \infty)$, but $(0, \infty)$ and $(-\infty, \infty)$ also arise.

In Section 6.10 we introduce the strong and weak versions of the M_2 topology, denoted by SM_2 and WM_2 . In Section 6.11 we provide alternative characterizations of these topologies and discuss additional properties.

Finally, in Section 6.12 we discuss characterizations of compact subsets of D using oscillation functions. These characterizations are useful because they lead to characterizations of tightness for sequences of probability measures on D , which is a principal way to establish weak convergence of the probability measures; see Section 11.6 of the book.

6.2. Regularity Properties of D

Recall that $D \equiv D^k \equiv D([0, T], \mathbb{R}^k)$ is the set of all $\mathbb{R}^k$ -valued functions $x \equiv (x^1, \dots, x^k)$ on $[0, T]$ that are right continuous at all $t \in [0, T)$ and have left limits at all $t \in (0, T]$:

We use superscripts to designate coordinate functions, so that subscripts can index different functions in D . For example, x_3^2 denotes the second coordinate function in $D([0, T], \mathbb{R}^1)$ of $x_3 \equiv (x_3^1, \dots, x_3^k)$ in $D([0, T], \mathbb{R}^k)$, where x_3 is the third element of the sequence $\{x_n : n \geq 1\}$. Let C be the subset of continuous functions in D .

Let $\|\cdot\|$ be the maximum (or l_∞) norm on $\mathbb{R}^k$ and the *uniform norm* on D ; i.e., for each $b \equiv (b^1, \dots, b^k) \in \mathbb{R}^k$, let

$$\|b\| \equiv \max_{1 \leq i \leq k} |b^i| \quad (2.1)$$

and, for each $x \equiv (x^1, \dots, x^k) \in D([0, T], \mathbb{R}^k)$, let

$$\|x\| \equiv \sup_{0 \leq t \leq T} \|x(t)\| = \sup_{0 \leq t \leq T} \max_{1 \leq i \leq k} |x^i(t)|. \quad (2.2)$$

The maximum norm on $\mathbb{R}^k$ in (2.1) is topologically equivalent to the l_p norm

$$\|b\|_p \equiv \left(\sum_{i=1}^k (b^i)^p \right)^{1/p}.$$

For $p = 2$, the l_p norm is the Euclidean (or l_2) norm. For $p = 1$, the l_p norm is the sum (or l_1) norm. The uniform norm on D induces the uniform metric on D .

We first discuss regularity properties of D due to the existence of limits. Let $Disc(x)$ be the set of discontinuities of x , i.e.,

$$Disc(x) \equiv \{t \in (0, T] : x(t-) \neq x(t)\} \quad (2.3)$$

and let $Disc(x, \epsilon)$ be the set of discontinuities of magnitude at least ϵ , i.e.,

$$Disc(x, \epsilon) \equiv \{t \in (0, T] : \|x(t-) - x(t)\| \geq \epsilon\}. \quad (2.4)$$

The following is a key regularity property of D .

Theorem 6.2.1. (the number of discontinuities of a given size) *For each $x \in D$ and $\epsilon > 0$, $Disc(x, \epsilon)$ is a finite subset of $[0, T]$.*

Proof. We will show that $Disc(x, \epsilon)$ being infinite contradicts the existence of limits from the left and right. If $Disc(x, \epsilon)$ were infinite, then there would exist $t \in [0, T]$ and a sequence $\{t_n : n \geq 1\}$ with $t_n \in Disc(x, \epsilon)$ for all n and $t_n \downarrow t$ or $t_n \uparrow t$ as $n \rightarrow \infty$. Suppose that $t_n \downarrow t$; the other case is treated in the same way. Since $t_n \in Disc(x, \epsilon)$, we must have $\|x(t_n-) - x(t_n)\| \geq \epsilon$ for all n . Hence, there must exist another sequence $\{t'_n : n \geq 1\}$ such that $t_n > t'_n > t_{n+1} > t'_{n+1} > t$ for all n and $\|x(t_n) - x(t'_n)\| > \epsilon/2$ for all n . However, that contradicts the existence of limits from the right at t . ■

Corollary 6.2.1. (the number of discontinuities) *For each $x \in D$, $Disc(x)$ is either finite or countably infinite.*

Proof. Note that

$$Disc(x) = \bigcup_{n=1}^{\infty} Disc(x, n^{-1}). \quad \blacksquare$$

We say that a function x in D is *piecewise-constant* if there are finitely many time points t_i such that $0 \equiv t_0 < t_1 < \cdots < t_{m-1} \leq t_m \equiv T$ and x is

constant on the intervals $[t_{i-1}, t_i)$, $1 \leq i \leq m-1$, and $[t_{m-1}, T]$. Let D_c be the subset of piecewise-constant functions in D . Let $v(x; A)$ be the *modulus of continuity* of the function x over the set A , defined by

$$v(x; A) \equiv \sup_{t_1, t_2 \in A} \{ \|x(t_1) - x(t_2)\| \} \quad (2.5)$$

for $A \subseteq [0, T]$. The following is a second important regularity property of D .

Theorem 6.2.2. (approximation by piecewise-constant functions) *For each $x \in D$ and $\epsilon > 0$, there exists $x_c \in D_c$ such that $\|x - x_c\| < \epsilon$.*

Proof. We show how to construct x_c . Given x and ϵ , construct the subset $Disc(x, \epsilon)$, which is finite by Theorem 6.2.1. Due to the existence of limits, for each $t \in Disc(x, \epsilon)$ we can find $t_1 \equiv t_1(t)$ and $t_2 \equiv t_2(t)$ such that $t_1 < t < t_2$, $v(x, [t_1, t]) < \epsilon$, $v(x, [t, t_2]) < \epsilon$,

$$Disc(x, \epsilon) \cap [t_1, t] = \emptyset \quad \text{and} \quad Disc(x, \epsilon) \cap (t, t_2] = \emptyset.$$

For each $t \in Disc(x, \epsilon)$, let these points t , $t_1(t)$ and $t_2(t)$ all belong to $Disc(x_c)$; let $x_c(t') = x(t-)$ for $t' \in (t_1, t)$ and let $x_c(t') = x(t)$ for $t' \in [t, t_2]$. Now let

$$A \equiv [0, T] - \bigcup_{t \in Disc(x, \epsilon)} (t_1(t), t_2(t)).$$

The set A is a finite union of closed intervals. Consider any one of these intervals, say $[a, b]$. If $v(x; [a, b]) < \epsilon$, then it suffices to let $x_c(t) = x(t)$ for any $t \in [a, b]$, and not add any points to $Disc(x_c)$. Suppose that $v(x; [a, b]) \geq \epsilon$. For each $t \in [a, b]$, since $t \in Disc(x, \epsilon)^c$, it is possible to find an interval $(t_1(t), t_2(t))$, $[a, t_2(t))$ or $(t_1(t), b]$ containing t such that $v(x, (t_1(t), t_2(t))) < \epsilon$. (The intervals $[a, t)$ and $(t, b]$ are open in the relative topology on $[a, b]$. Thus the collection of all these subintervals form an open cover of $[a, b]$.) Since $[a, b]$ is compact, there is a finite collection of these intervals covering $[a, b]$; i.e., there are points

$$a < t'_1 < t_1 < \cdots < t'_m < t_m < b$$

for $m \geq 1$ such that $[a, t_1)$, (t'_1, t_2) , (t'_2, t_3) , $\dots$, (t'_{m-1}, t_m) , $(t'_m, b]$ are in the finite collection. Necessarily, $t'_i < t_i$ for all i . It suffices to choose $t''_i \in (t'_i, t_i)$ for each i , $1 \leq i \leq m$, and let $t''_i \in Disc(x_c)$. We can let $x_c(t''_i) = x(t''_i)$ for each such t''_i . We have thus constructed $x_c \in D_c$ with $\|x - x_c\| < \epsilon$. ■

6.3. Strong and Weak M_1 Topologies

6.3.1. Definitions

We start by making some definitions, repeating what is in the book. The strong and weak topologies will be based on different notions of a segment in $\mathbb{R}^k$. For $a \equiv (a^1, \dots, a^k)$, $b \equiv (b^1, \dots, b^k) \in \mathbb{R}^k$, let $[a, b]$ be the *standard segment*, i.e.,

$$[a, b] \equiv \{\alpha a + (1 - \alpha)b : 0 \leq \alpha \leq 1\} \quad (3.1)$$

and let $[[a, b]]$ be the *product segment*, i.e.,

$$[[a, b]] \equiv \bigtimes_{i=1}^k [a^i, b^i] \equiv [a^1, b^1] \times \dots \times [a^k, b^k], \quad (3.2)$$

where the one-dimensional segment $[a^i, b^i]$ coincides with the closed interval $[a^i \wedge b^i, a^i \vee b^i]$, with $c \wedge d = \min\{c, d\}$ and $c \vee d = \max\{c, d\}$ for $c, d \in \mathbb{R}$. Note that $[a, b]$ and $[[a, b]]$ are both subsets of $\mathbb{R}^k$. If $a = b$, then $[a, b] = [[a, b]] = \{a\} = \{b\}$; if $a^i \neq b^i$ for one and only one i , then $[a, b] = [[a, b]]$. If $a \neq b$, then $[a, b]$ is always a one-dimensional line in $\mathbb{R}^k$, while $[[a, b]]$ is a j -dimensional subset, where j is the number of coordinates i for which $a^i \neq b^i$. Always, $[a, b] \subseteq [[a, b]]$.

We now define completed graphs of the functions: For $x \in D$, let the (standard) *thin graph* of x be

$$\Gamma_x \equiv \{(z, t) \in \mathbb{R}^k \times [0, T] : z \in [x(t-), x(t)]\}, \quad (3.3)$$

where $x(0-) \equiv x(0)$ and let the *thick graph* of x be

$$\begin{aligned} G_x &\equiv \{(z, t) \in \mathbb{R}^k \times [0, T] : z \in [[x(t-), x(t)]]\} \\ &= \{(z, t) \in \mathbb{R}^k \times [0, T] : z^i \in [x^i(t-), x^i(t)] \text{ for each } i\} \end{aligned} \quad (3.4)$$

for $1 \leq i \leq k$. Since $[a, b] \subseteq [[a, b]]$ for all $a, b \in \mathbb{R}^k$, $\Gamma_x \subseteq G_x$ for each x .

We now define *order relations* on the graphs Γ_x and G_x . We say that $(z_1, t_1) \leq (z_2, t_2)$ if either (i) $t_1 < t_2$ or (ii) $t_1 = t_2$ and $|x^i(t_1-) - z_1^i| \leq |x^i(t_1-) - z_2^i|$ for all i . The relation $\leq$ induces a total order on Γ_x and a partial order on G_x .

It is also convenient to look at the ranges of the functions. Let the *thin range* of x be the projection of Γ_x onto $\mathbb{R}^k$, i.e.,

$$\rho(\Gamma_x) \equiv \{z \in \mathbb{R}^k : (z, t) \in \Gamma_x \text{ for some } t \in [0, T]\} \quad (3.5)$$

and let the *thick range* of x be the projection of G_x onto $\mathbb{R}^k$, i.e.,

$$\rho(G_x) \equiv \{z \in \mathbb{R}^k : (z, t) \in G_x \text{ for some } t \in [0, T]\} . \quad (3.6)$$

Note that $(z, t) \in \Gamma_x(G_x)$ for some t if and only if $z \in \rho(\Gamma_x)$ ($\rho(G_x)$). Thus a pair (z, t) cannot be in a graph of x if z is not in the corresponding range.

We now define strong (standard) and weak parametric representations based on these two kinds of graphs. A *strong parametric representation* of x is a continuous nondecreasing function (u, r) mapping $[0, 1]$ onto Γ_x . A *weak parametric representation* of x is a continuous nondecreasing function (u, r) mapping $[0, 1]$ into G_x such that $r(0) = 0$, $r(1) = T$ and $u(1) = x(T)$. (For the parametric representation, “nondecreasing” is with respect to the usual order on the domain $[0, 1]$ and the order on the graphs defined above.) Here it is understood that $u \equiv (u^1, \dots, u^k) \in C([0, 1], \mathbb{R}^k)$ is the spatial part of the parametric representation, while $r \in C([0, 1], [0, T])$ is the time (domain) part. Let $\Pi_s(x)$ and $\Pi_w(x)$ be the sets of strong and weak parametric representations of x , respectively. For real-valued functions x , let $\Pi(x) \equiv \Pi_s(x) = \Pi_w(x)$. Note that $(u, r) \in \Pi_w(x)$ if and only if $(u^i, r) \in \Pi(x^i)$ for $1 \leq i \leq k$.

We use the parametric representations to characterize the strong and weak M_1 topologies. As in (2.1) and (2.2), let $\|\cdot\|$ denote the supremum norms in $\mathbb{R}^k$ and D . We use the definition $\|\cdot\|$ in (2.2) also for the $\mathbb{R}^k$ -valued functions u and r on $[0, 1]$.

Now, for any $x_1, x_2 \in D$, let

$$d_s(x_1, x_2) \equiv \inf_{\substack{(u_j, r_j) \in \Pi_s(x_j) \\ j=1,2}} \{\|u_1 - u_2\| \vee \|r_1 - r_2\|\} \quad (3.7)$$

and

$$d_w(x_1, x_2) \equiv \inf_{\substack{(u_j, r_j) \in \Pi_w(x_j) \\ j=1,2}} \{\|u_1 - u_2\| \vee \|r_1 - r_2\|\} . \quad (3.8)$$

Note that $\|u_1 - u_2\| \vee \|r_1 - r_2\|$ can also be written as $\|(u_1, r_1) - (u_2, r_2)\|$, due to definitions (2.1) and (2.2). Of course, when the range is $\mathbb{R}$, $d_s = d_w = d_{M_1}$ for d_{M_1} defined in equation (3.4) in Section 3.3 of the book.

We say that $x_n \rightarrow x$ in D for a sequence or net $\{x_n\}$ in the $SM_1(WM_1)$ topology if $d_s(x_n, x) \rightarrow 0$ ($d_w(x_n, x) \rightarrow 0$) as $n \rightarrow \infty$. We start with the following basic result.

6.3.2. Metric Properties

Theorem 6.3.1. (metric inducing SM_1) d_s is a metric on D .

Proof. Only the triangle inequality is difficult. By Lemma 6.3.2 below, for any $\epsilon > 0$, a common parametric representation $(u_3, r_3) \in \Pi_s(x_3)$ can be used to obtain

$$\|u_1 - u_3\| \vee \|r_1 - r_3\| < d_s(x_1, x_3) + \epsilon$$

and

$$\|u_2 - u_3\| \vee \|r_2 - r_3\| < d_s(x_2, x_3) + \epsilon$$

for some $(u_1, r_1) \in \Pi_s(x_1)$ and $(u_2, r_2) \in \Pi_s(x_2)$. Hence

$$d_s(x_1, x_2) \leq \|u_1 - u_2\| \vee \|r_1 - r_2\| \leq d_s(x_1, x_3) + d_s(x_3, x_2) + 2\epsilon .$$

Since ϵ was arbitrary, the proof is complete. ■

To prove Theorem 6.3.1, we use finite approximations to the graphs Γ_x . We first define an order-consistent distance between a graph and a finite subset. We use the notion of a finite ordered subset.

Definition 6.3.1. (order-consistent distance) *For $x \in D$, let A be a finite ordered subset of the ordered graph $(\Gamma_x, \leq)$, i.e., for some $m \geq 1$, A contains $m + 1$ points (z_i, t_i) from Γ_x such that*

$$(x(0), 0) \equiv (z_0, t_0) \leq (z_1, t_1) \leq \cdots \leq (z_m, t_m) \equiv (x(T), T) . \quad (3.9)$$

The order-consistent distance between A and Γ_x is

$$\hat{d}(A, \Gamma_x) \equiv \sup\{\|(z, t) - (z_i, t_i)\| \vee \|(z, t) - (z_{i+1}, t_{i+1})\|\} , \quad (3.10)$$

where the supremum is over all $(z_i, t_i) \in A$, $1 \leq i \leq m-1$, and all $(z, t) \in \Gamma_x$ such that

$$(z_i, t_i) \leq (z, t) < (z_{i+1}, t_{i+1}) ,$$

using the order on the graph. ■

We now show that finite ordered subsets A can be chosen to make $\hat{d}(A, \Gamma_x)$ arbitrarily small.

Lemma 6.3.1. (finite approximations to graphs) *For any $x \in D$ and $\epsilon > 0$, there exists a finite ordered subset A of Γ_x such that $\hat{d}(A, \Gamma_x) < \epsilon$ for $\hat{d}$ in (3.10).*

Proof. First put finitely many points $(x(t_i), t_i)$ in A to meet the requirement on the domain $[0, T]$, i.e., to have $0 = t_1 < t_2 < \cdots < t_m = T$ with $t_{i+1} - t_i < \epsilon$. We add additional points to account for the spatial component. For each $t \in \text{Disc}(x, \epsilon)$, choose the points $(x(t-), t)$, $(x(t), t)$ and finitely many points on the segment $[(x(t-), t), (x(t), t)]$ such that the distance between successive points is less than ϵ . Since x has left and right limits everywhere, there are open neighborhoods (t_1, t) and (t, t_2) of each $t \in \text{Disc}(x, \epsilon)$ such that

$$\sup\{\|x(t') - x(t'')\| : t_1 < t' < t'' < t\} < \epsilon$$

and

$$\sup\{\|x(t') - x(t'')\| : t < t' < t'' < t_2\} < \epsilon.$$

We thus can choose one more point, if needed, in each of the sets $\Gamma_x \cap [R^k \times (t_1, t)]$ and $\Gamma_x \cap [R^k \times (t, t_2)]$ to achieve the desired property over each open interval (t_1, t_2) in $[0, T]$. The complement of the union of these finitely many open intervals in $[0, T]$ is a compact subset of $[0, T]$. Knowing that (i) all remaining discontinuities are of magnitude less than ϵ and (ii) limits exist everywhere from the left and right, we can conclude that there is a closed interval of positive length about each point in the compact set, where x oscillates by less than ϵ , i.e., $\sup\{\|x(t') - x(t'')\| < \epsilon$, where t', t'' are points in the interval. However, by the compactness, only finitely many of these closed intervals cover the compact set. We add points $(x(t), t)$ to A to ensure that there is at least one point (z, t) for which t is in one of these closed intervals. By this construction, A is finite and $\hat{d}(A, \Gamma_x) < \epsilon$. ■

To complete the proof of Theorem 6.3.1, we need the following result, which we prove by applying Lemma 6.3.1.

Lemma 6.3.2. (flexibility in choice of parametric representations) *For any $x_1, x_2 \in D$, $(u_1, r_1) \in \Pi_s(x_1)$ and $\epsilon > 0$, it is possible to find $(u_2, r_2) \in \Pi_s(x_2)$ such that*

$$\|u_1 - u_2\| \vee \|r_1 - r_2\| \leq d_s(x_1, x_2) + \epsilon.$$

Proof. For $x_1, x_2 \in D$ and ϵ given, choose $(u'_1, r'_1) \in \Pi_s(x_1)$ and $(u'_2, r'_2) \in \Pi_s(x_2)$ such that

$$\|u'_1 - u'_2\| \vee \|r'_1 - r'_2\| < d_s(x_1, x_2) + \epsilon/4. \quad (3.11)$$

Next apply Lemma 6.3.1 to find a finite ordered subsets $A_1 \subseteq \Gamma_{x_1}$ such that $\hat{d}(A_1, \Gamma_{x_1}) < \epsilon/4$. Next find a finite subset S'_1 of $[0, 1]$ of the same cardinality

as A_1 such that $(u'_1(s), r'_1(s)) \in A_1$ for each $s \in S'_1$. Let S_1 be another finite subset of $[0, 1]$ of the same cardinality as A_1 such that $(u_1(s), r_1(s)) \in A_1$ for each $s \in S_1$. Let λ be a homeomorphism of $[0, 1]$ such that λ maps S_1 onto S'_1 . Let $(u_2, r_2) = (u'_2 \circ \lambda, r'_2 \circ \lambda)$, where $\circ$ is the composition map. Trivially, by (3.11),

$$\|u'_1 \circ \lambda - u'_2 \circ \lambda\| \vee \|r'_1 \circ \lambda - r'_2 \circ \lambda\| < d_s(x_1, x_2) + \epsilon/4 .$$

Hence, it suffices to show that

$$\|u_1 - u'_1 \circ \lambda\| \vee \|r_1 - r'_1 \circ \lambda\| < 3\epsilon/4 . \quad (3.12)$$

First there is equality $u_1(s) = u'_1(\lambda(s))$ by construction at each $s \in S_1$. However, since $\hat{d}(A_1, \Gamma_x) < \epsilon/4$, (3.12) holds: For each $s \in [0, 1]$, there is $s_i \in S_1$ such that $s_i \leq s < s_{i+1}$ and

$$\begin{aligned} \|u_1(s) - u'_1(\lambda(s))\| &\leq \|u_1(s) - u_1(s_i)\| + \|u_1(s_i) - u'_1(\lambda(s_i))\| \\ &\quad + \|u'_1(\lambda(s_i)) - u'_1(\lambda(s))\| \leq \epsilon/2 . \quad \blacksquare \end{aligned}$$

We will show that the metric d_s induces the standard M_1 topology defined by Skorohod (1956); see Theorem 6.5.1. Since $\Pi_s(x) \subseteq \Pi_w(x)$ for all x , we have $d_w(x_1, x_2) \leq d_s(x_1, x_2)$ for all x_1, x_2 , so that the WM_1 topology is indeed weaker than the SM_1 topology. However, we show below in Example 12.3.2 of the book that d_w in (3.8) is *not* a metric when $k > 1$.

For $x_1, x_2 \in D([0, T], \mathbb{R}^k)$, let d_p be a metric inducing the product topology, defined by

$$d_p(x_1, x_2) \equiv \max_{1 \leq i \leq k} d(x_1^i, x_2^i) \quad (3.13)$$

for $x_j \equiv (x_j^1, \dots, x_j^k)$ and $j = 1, 2$. (Note that $d_s = d_w = d_p$ when the functions are real valued, in which case we use the notation d .) It is an easy consequence of (3.8), (3.13) and the second representation in (3.4) that the WM_1 topology is stronger than the product topology, i.e., $d_p(x_1, x_2) \leq d_w(x_1, x_2)$ for all $x_1, x_2 \in D$. In Section 6.5 we will show that actually the WM_1 and product topologies coincide.

Example 12.3.1 of the book shows that SM_1 is strictly stronger than WM_1 .

We now relate the metrics $d_{M_1} \equiv d_s$ and d_{J_1} for d_{J_1} in equation 3.2 of Section 3.3 in the book.

Theorem 6.3.2. (comparison of J_1 and M_1 metrics) *For each $x_1, x_2 \in D$,*

$$d_s(x_1, x_2) \leq d_{J_1}(x_1, x_2) .$$

Proof. For any $x_1, x_2 \in D$ and $\lambda \in \Lambda$, we show how to define parametric representations (u_j, r_j) in $\Pi_s(x_j)$ for $j = 1, 2$ such that

$$\|u_1 - u_2\| \vee \|r_1 - r_2\| = \|x_1 \circ \lambda - x_2\| \vee \|\lambda - e\|. \quad (3.14)$$

If, for any $\epsilon > 0$, we first choose $\lambda \in \Lambda$ so that

$$\|x_1 \circ \lambda - x_2\| \vee \|\lambda - e\| \leq d_{J_1}(x_1, x_2) + \epsilon,$$

the associated parametric representation yield

$$d_s(x_1, x_2) \leq \|u_1 - u_2\| \vee \|r_1 - r_2\| \leq d_{J_1}(x_1, x_2) + \epsilon.$$

Since ϵ is arbitrary, that will complete the proof. Suppose that

$$t_n \in \text{Disc}(x_1, x_2) \equiv \text{Disc}(x_1) \cup \text{Disc}(x_2), \quad n \geq 1,$$

where t_n is ordered (indexed) first by the norm of the jump and then the location, with values closer to 0 occurring first. Associate with each time point t_n a closed subinterval $[a_n, b_n]$ in $(0, 1)$ such that the subintervals are ordered, i.e., if $t_i < t_j < t_k$ are three points in $\text{Disc}(x_1, x_2)$, then $a_i < b_i < a_j < b_j < a_k < b_k$. Then let $r_2(s) = t_n$ for $a_n \leq s \leq b_n$. If $t \notin \text{Disc}(x_1, x_2)$ but $t_{n_k} \downarrow t$ as $n_k \rightarrow \infty$ for $t_{n_k} \in \text{Disc}(x_1, x_2)$, then let $r_2(s) = \lim_{n_k \rightarrow \infty} r_2(a_{n_k})$. Similarly, if $t \notin \text{Disc}(x_1, x_2)$ but $t_{n_k} \uparrow t$ as $n_k \rightarrow \infty$ for $t_{n_k} \in \text{Disc}(x_1, x_2)$, then let $r_2(s) = \lim_{n_k \rightarrow \infty} r_2(b_{n_k})$. Finally, let $r_2(s)$ be defined by linear interpolation in all remaining gaps. This makes r_2 continuous and nondecreasing. Having defined r_2 , let $r_1 = \lambda \circ r_2$, $u_1(s) = (x_1 \circ r_1)(s)$ and $u_2(s) = (x_2 \circ r_2)(s)$ for all s , except $s \in (a_n, b_n)$ for some n . Within each subinterval (a_n, b_n) , let u_1 and u_2 be defined by linear interpolation from their values at the endpoints a_n and b_n . This construction makes $(u_j, r_j) \in \Pi_s(x_j)$ for $j = 1, 2$ and yields (3.14), thus completing the proof. ■

6.3.3. Properties of Parametric Representations

We conclude this section by further discussing strong parametric representations. For $x \in D$, $t \in \text{Disc}(x)$ and $(u, r) \in \Pi_s(x)$, there exists a unique pair of points $s_- \equiv s_-(t, x)$ and $s_+ \equiv s_+(t, x)$ such that $s_- < s_+$ and $r^{-1}(\{t\}) = [s_-, s_+]$, i.e.,

- (i) $r(s) < t$ for $s < s_-$
- (ii) $r(s) = t$ for $s_- \leq s \leq s_+$
- (iii) $r(s) > t$ for $s > s_+$.

(3.15)

We will exploit the fact that a parametric representation (u, r) in $\Pi_s(x)$ is *jump consistent*: for each $t \in \text{Disc}(x)$ and pair $s_- \equiv s_-(t, x) < s_+ \equiv s_+(t, x)$ such that (3.15) holds, there is a continuous nondecreasing function β_t mapping $[0, 1]$ onto $[0, 1]$ such that

$$u(s) = \beta_t \left(\frac{s - s_-}{s_+ - s_-} \right) u(s_+) + \left[1 - \beta_t \left(\frac{s - s_-}{s_+ - s_-} \right) \right] u(s_-) \quad \text{for } s_- \leq s \leq s_+ . \quad (3.16)$$

Condition (3.16) means that u is defined within jumps by interpolation from the definition at the endpoints s_- and s_+ , consistently over all coordinates. In particular, suppose that $t \in \text{Disc}(x^i)$. (Since $t \in \text{Disc}(x)$, we must have $t \in \text{Disc}(x^i)$ for some coordinate i .) Suppose that $x^i(t-) < x^i(t)$. Then we can let

$$\beta_t(s) = \frac{u^i(s) - u^i(s_-)}{u^i(s_+) - u^i(s_-)} . \quad (3.17)$$

We see that (3.16) and (3.17) are consistent in that

$$u^i(s) = \beta_t \left(\frac{s - s_-}{s_+ - s_-} \right) u^i(s_+) + \left[1 - \beta_t \left(\frac{s - s_-}{s_+ - s_-} \right) \right] u^i(s_-) \quad (3.18)$$

for β_t in (3.17). For another coordinate j , (3.16) and (3.17) imply that

$$u^j(s) = \left(\frac{u^i(s) - u^i(s_-)}{u^i(s_+) - u^i(s_-)} \right) u^j(s_+) + \left(\frac{u^i(s_+) - u^i(s)}{u^i(s_+) - u^i(s_-)} \right) u^j(s_-) . \quad (3.19)$$

It is possible that $t \notin \text{Disc}(x^j)$, in which case $u^j(s) = u^j(s_-) = u^j(s_+)$ for all s , $s_- \leq s \leq s_+$.

We can further characterize the behavior of a strong parametric representation at a discontinuity point. For $x \in D$, $t \in \text{Disc}(x)$ and $(u, r) \in \Pi_s(x)$, there exists a unique set of four points $s_- \equiv s_-(t, x) \leq s'_- \equiv s'_-(t, x) < s'_+ \equiv s'_+(t, x) \leq s_+ \equiv s_+(t, x)$ such that (3.15) holds and

$$\begin{aligned} & \text{(i) } u(s) = u(s_-) \text{ for } s_- \leq s \leq s'_-, \\ & \text{(ii) for each } i, \text{ either } u^i(s_-) < u^i(s) < u^i(s_+), \\ & \quad \text{or } u^i(s_-) > u^i(s) > u^i(s_+) \text{ for } s'_- < s < s'_+, \\ & \text{(iii) } u(s) = u(s_+) \text{ for } s'_+ \leq s \leq s_+ . \end{aligned} \quad (3.20)$$

Let D_1 be the subset of D containing functions all of whose jumps occur in only one coordinate, i.e., the set of x such that, for each $t \in \text{Disc}(x)$ there exists one and only one $i \equiv i(t)$ such that $t \in \text{Disc}(x^i)$. (The coordinate i may depend on t .)

Lemma 6.3.3. (strong and weak parametric representations coincide on D_1) For each $x \in D_1$, $\Pi_s(x) = \Pi_w(x)$.

Proof. Since $\Pi_s(x) \subseteq \Pi_w(x)$, we need to show that $(u, r) \in \Pi_w(x)$ is in $\Pi_s(x)$ for x in $D^{(1)}$. Pick any $t \in \text{Disc}(x)$ and let i be the coordinate of x with a jump at t . We can then define the β_t needed for (3.16) using (3.17). Since $u^j(s) = u^j(s_-) = u^j(s_+)$ for all j with $j \neq i$, (3.19) and (3.16) are then satisfied. ■

Corollary. For each $x \in D([0, T], \mathbb{R}^1)$, $\Pi_s(x) = \Pi_w(x)$.

We now show that parametric representations are preserved under linear functions of the coordinates when $x \in \Pi_s(x)$. That is *not* true in $\Pi_w(x)$.

Lemma 6.3.4. (linear functions of parametric representations) If $(u, r) \in \Pi_s(x)$, then $(\eta u, r) \in \Pi_s(\eta x)$ for any $\eta \in \mathbb{R}^k$.

Proof. By the Corollary to Lemma 6.3.3, $\Pi_s(\eta x) = \Pi_w(\eta x)$. Hence, it suffices to show that $(\eta u, r) \in \Pi_w(\eta x)$. It is clear that $(\eta u, r)$ is continuous and nondecreasing. For $t \in \text{Disc}(\eta x)$, necessarily $t \in \text{Disc}(x)$. (We could have $t \in \text{Disc}(x)$ but $t \notin \text{Disc}(\eta x)$, but that does not concern us.) By (3.16), when $r(s) = t$,

$$\eta u(s) = \beta_t \left(\frac{s - s_-}{s_+ - s_-} \right) \eta u(s_+) + \left[1 - \beta_t \left(\frac{s - s_-}{s_+ - s_-} \right) \right] \eta u(s_-)$$

which completes the proof. ■

6.4. Local Uniform Convergence at Continuity Points

In this section we provide alternative characterizations of local uniform convergence at continuity points of a limit function. The non-uniform Skorohod topologies on D all imply local uniform convergence at continuity points of a limit function. They differ by their behavior at discontinuity points.

We start by defining two basic *uniform-distance functions*. For $x_1, x_2 \in D$, $t \in [0, T]$ and $\delta > 0$, let

$$u(x_1, x_2, t, \delta) \equiv \sup_{0 \vee (t-\delta) \leq t_1 \leq (t+\delta) \wedge T} \{ \|x_1(t_1) - x_2(t_1)\| \}, \quad (4.1)$$

$$v(x_1, x_2, t, \delta) \equiv \sup_{0 \vee (t-\delta) \leq t_1, t_2 \leq (t+\delta) \wedge T} \{ \|x_1(t_1) - x_2(t_2)\| \}, \quad (4.2)$$

We also define an *oscillation function*. For $x \in D$, $t \in [0, T]$ and $\delta > 0$, let

$$\bar{v}(x, t, \delta) \equiv \sup_{0 \vee (t-\delta) \leq t_1 \leq t_2 \leq (t+\delta) \wedge T} \{\|x(t_1) - x(t_2)\|\} . \quad (4.3)$$

We next define oscillation functions that we will use with the M_1 topologies. They use the distance $\|z - A\|$ between a point z and a subset A in $\mathbb{R}^k$ defined in equation 5.3 in Section 11.5 of the book. The SM_1 and WM_1 topologies use the standard and product segments in (3.1) and (3.2). For each $x \in D$, $t \in [0, T]$ and $\delta > 0$, let

$$w_s(x, t, \delta) \equiv \sup_{0 \vee (t-\delta) \leq t_1 < t_2 < t_3 \leq (t+\delta) \wedge T} \{\|x(t_2) - [x(t_1), x(t_3)]\|\} \quad (4.4)$$

and

$$w_w(x, t, \delta) \equiv \sup_{0 \vee (t-\delta) \leq t_1 < t_2 < t_3 \leq (t+\delta) \wedge T} \{\|x(t_2) - [[x(t_1), x(t_3)]]\|\} \quad (4.5)$$

We now turn to the M_2 topology, which we will be studying in Sections 6.10 and 6.11. We define two uniform-distance functions. We use $\bar{w}$ as opposed to w to denote an M_2 uniform-distance function. Just as with the M_1 topologies, the SM_2 and WM_2 topologies use the standard and product segments in (3.1) and (3.2). For $x_1, x_2 \in D$, let

$$\bar{w}_s(x_1, x_2, t, \delta) \equiv \sup_{0 \vee (t-\delta) \leq t_1 \leq (t+\delta) \wedge T} \{\|x_1(t_1) - [x_2(t-), x_2(t)]\|\} \quad (4.6)$$

$$\bar{w}_w(x_1, x_2, t, \delta) \equiv \sup_{0 \vee (t-\delta) \leq t_1 \leq (t+\delta) \wedge T} \{\|x_1(t_1) - [[x_2(t-), x_2(t)]]\|\} \quad (4.7)$$

It is easy to establish the following relations among the uniform-distance and oscillation functions.

Lemma 6.4.1. (inequalities for uniform-distance and oscillation functions)
For all $x, x_n \in D$, $t \in [0, T]$ and $\delta > 0$,

$$u(x_n, x, t, \delta) \leq v(x_n, x, t, \delta) \leq u(x_n, x, t, \delta) + \bar{v}(x, t, \delta) ,$$

$$w_w(x_n, t, \delta) \leq w_s(x_n, t, \delta) \leq \bar{v}(x_n, t, \delta) \leq 2v(x_n, x, t, \delta) + \bar{v}(x, t, \delta) ,$$

$$\bar{w}_w(x_n, x, t, \delta) \leq \bar{w}_s(x_n, x, t, \delta) \leq v(x_n, x, t, \delta) \leq 2\bar{w}_w(x_n, x, t, \delta) + \bar{v}(x, t, \delta) .$$

Since the M_1 -oscillation functions $w_s(x_n, t, \delta)$ and $w_w(x_n, t, \delta)$ do not contain the limit x , their convergence to 0 as $n \rightarrow \infty$ and then $\delta \downarrow 0$ does not directly imply local uniform convergence at a continuity point of a prospective limit function x .

We relate convergence of $w_s(x_n, t, \delta)$ and $w_w(x_n, t, \delta)$ to 0 as $n \rightarrow \infty$ and $\delta \downarrow 0$ to local uniform convergence by requiring pointwise convergence in a neighborhood of t ; see (vi) in Theorem 6.4.1 below.

Theorem 6.4.1. (characterizations of local uniform convergence at continuity points) *If $t \notin \text{Disc}(x)$, then the following are equivalent:*

$$(i) \quad \lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} u(x_n, x, t, \delta) = 0, \quad (4.8)$$

$$(ii) \quad \lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} v(x_n, x, t, \delta) = 0, \quad (4.9)$$

$$(iii) \quad \lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{w}_s(x_n, x, t, \delta) = 0, \quad (4.10)$$

$$(iv) \quad \lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{w}_w(x_n, x, t, \delta) = 0, \quad (4.11)$$

(v) $x_n(t_1) \rightarrow x(t_1)$ for all t_1 in a dense subset of a neighborhood of t (including 0 if $t = 0$ or T if $t = T$) and

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w_s(x_n, t, \delta) = 0,$$

(vi) $x_n(t_1) \rightarrow x(t_1)$ for all t_1 in a dense subset of a neighborhood of t (including 0 if $t = 0$ or T if $t = T$) and

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w_w(x_n, t, \delta) = 0. \quad (4.12)$$

Proof. By Lemma 6.4.1, we have the implications (i) $\leftrightarrow$ (ii) $\leftrightarrow$ (iii) $\leftrightarrow$ (iv) and (ii) $\rightarrow$ (v) $\rightarrow$ (vi). Hence it suffices to show that (vi) $\rightarrow$ (i), which we now do. For $x, t \notin \text{Disc}(x)$ and $\epsilon > 0$ given, choose $\delta > 0$ so that $\bar{v}(x, t, \delta) < \epsilon$, which is possible since $t \notin \text{Disc}(x)$. Also let δ be sufficiently small so that $x_n(t'_1) \rightarrow x(t'_1)$ as $n \rightarrow \infty$ for all t'_1 in a dense subset of $[0 \vee (t - \delta), (t + \delta) \wedge T]$. Note that we can treat 0 and T directly. For $t_1 \in (0 \vee (t - \delta), T \wedge (t + \delta))$ given, choose t'_1, t'_2 so that $0 \vee (t - \delta) < t'_1 < t_1 < t'_2 < (t + \delta) \wedge T$ and $x_n(t'_j) \rightarrow x(t'_j)$

as $n \rightarrow \infty$ for $j = 1, 2$. Then choose n_0 so that $\|x_n(t') - x(t'')\| < \epsilon$ for $t'' = 0, T, t'_1$ and t'_2 and $w_w(x_n, t, \delta) < \epsilon$ for $n \geq n_0$. Then, for $n \geq n_0$,

$$\begin{aligned}
\|x_n(t_1) - x(t_1)\| &\leq \|x_n(t_1) - x_n(t'_1)\| + \|x_n(t'_1) - x(t'_1)\| + \|x(t'_1) - x(t_1)\| \\
&\leq \|x_n(t_1) - x_n(t'_1)\| + 2\epsilon \\
&\leq \|x_n(t_1) - [x_n(t'_1), x_n(t'_2)]\| + \|x_n(t'_1) - x_n(t'_2)\| + 2\epsilon \\
&\leq w_w(x_n, t, \delta) + \|x_n(t'_1) - x_n(t'_2)\| + 2\epsilon \\
&\leq \|x_n(t'_1) - x(t'_1)\| + \|x(t'_1) - x(t'_2)\| \\
&\quad + \|x(t'_2) - x_n(t'_2)\| + 3\epsilon \leq 6\epsilon .
\end{aligned}$$

It remains to consider $t = 0$ and $t = T$. The reasoning is the same for these two cases, so we consider only $t = 0$. For $t = 0$, note that

$$\|x_n(t_1) - x(t_1)\| \leq \|x_n(t_1) - x_n(0)\| + \|x_n(0) - x(0)\| + \|x(0) - x(t)\| . \quad (4.13)$$

The third term in (4.13) can be made small using the right continuity of x at 0; the second term in (4.13) can be made small by the assumed convergence at 0; the first term in (4.13) can be made small by (4.12). ■

We now show that local uniform convergence at all points in a compact interval implies uniform convergence over the compact interval.

Lemma 6.4.2. (local uniform convergence everywhere in a compact interval) *If (4.8) holds for all $t \in [a, b]$, then*

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \sup_{0 \vee (a-\delta) \leq t \leq (b+\delta) \wedge T} \{\|x_n(t) - x(t)\|\} = 0 .$$

Proof. By (4.8), for all $\epsilon > 0$ and $t \in [a, b]$, there exists $\delta(t)$ such that

$$\overline{\lim}_{n \rightarrow \infty} u(x_n, x, t, \delta(t)) < \epsilon .$$

For each t , there is thus uniform asymptotic closeness in the intervals $(0 \vee (t - \delta(t)), (t + \delta(t)) \wedge T)$. However, these intervals form an open cover of the interval $[a, b]$. Since $[a, b]$ is compact, there is a finite subcover. Hence, there is a $\delta' > 0$ such that

$$\overline{\lim}_{n \rightarrow \infty} \sup_{0 \vee (a-\delta') \leq t \leq (b+\delta') \wedge T} \{\|x_n(t) - x(t)\|\} < \epsilon .$$

Since ϵ was arbitrary, this implies the desired conclusion. ■

6.5. Alternative Characterizations of M_1 Convergence

We now give alternative characterizations of SM_1 and WM_1 convergence.

6.5.1. SM_1 Convergence

We first establish alternative characterizations of SM_1 convergence or, equivalently, d_s -convergence. One characterization is a minor variant of the original one involving an oscillation function established by Skorohod (1956). Another one – (v) below – involves only the local behavior of the functions. It helps us establish sufficient conditions to have $d_s((x_n, y_n), (x, y)) \rightarrow 0$ in $D([0, T], \mathbb{R}^{k+l})$ when $d_s(x_n, x) \rightarrow 0$ in $D([0, T], \mathbb{R}^k)$ and $d_s(y_n, y) \rightarrow 0$ in $D([0, T], \mathbb{R}^l)$; see Section 6.6. For the SM_1 topology, we define another oscillation function. For any $x_1, x_2 \in D$ and $\delta > 0$, let

$$w_s(x, \delta) \equiv \sup_{0 \leq t \leq T} w_s(x, t, \delta) , \quad (5.1)$$

for $w_s(x, t, \delta)$ in (4.4).

The following main result is proved in the book. It only remains to prove the supporting lemmas, which we do here.

Theorem 6.5.1. (characterizations of SM_1 convergence) *The following are equivalent characterizations of convergence $x_n \rightarrow x$ as $n \rightarrow \infty$ in (D, SM_1) :*

(i) *For any $(u, r) \in \Pi_s(x)$, there exists $(u_n, r_n) \in \Pi_s(x_n)$, $n \geq 1$, such that*

$$\|u_n - u\| \vee \|r_n - r\| \rightarrow 0 \quad \text{as } n \rightarrow \infty . \quad (5.2)$$

(ii) *There exist $(u, r) \in \Pi_s(x)$ and $(u_n, r_n) \in \Pi_s(x_n)$ for $n \geq 1$ such that (5.2) holds.*

(iii) *$d_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$; i.e., for all $\epsilon > 0$ and all sufficiently large n , there exist $(u, r) \in \Pi_s(x)$ and $(u_n, r_n) \in \Pi_s(x_n)$ such that*

$$\|u_n - u\| \vee \|r_n - r\| < \epsilon .$$

(iv) *$x_n(t) \rightarrow x(t)$ as $n \rightarrow \infty$ for each t in a dense subset of $[0, T]$ including 0 and T , and*

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w_s(x_n, \delta) = 0 \quad (5.3)$$

for $w_s(x, \delta)$ in (5.1) and $w_s(x, t, \delta)$ in (4.4).

(v) $x_n(T) \rightarrow x(T)$ as $n \rightarrow \infty$; for each $t \notin \text{Disc}(x)$,

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} v(x_n, x, t, \delta) = 0 \quad (5.4)$$

for $v(x_1, x_2, t, \delta)$ in (4.2); and, for each $t \in \text{Disc}(x)$,

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w_s(x_n, t, \delta) = 0 \quad (5.5)$$

for $w_s(x, t, \delta)$ in (4.4).

(vi) For all $\epsilon > 0$, there exist integers m and n_1 , a finite ordered subset A of Γ_x of cardinality m as in (3.9) and, for all $n \geq n_1$, finite ordered subsets A_n of Γ_{x_n} of cardinality m such that, for all $n \geq n_1$, $\hat{d}(A, \Gamma_x) < \epsilon$, $\hat{d}(A_n, \Gamma_{x_n}) < \epsilon$ for $\hat{d}$ in (3.10) and $d^*(A, A_n) < \epsilon$, where

$$d^*(A, A_n) \equiv \max_{1 \leq i \leq m} \{ \| (z_i, t_i) - (z_{n,i}, t_{n,i}) \| : (z_i, t_i) \in A, (z_{n,i}, t_{n,i}) \in A_n \}. \quad (5.6)$$

In preparation for the proof of Theorem 6.5.1, we establish some preliminary results. We first show that SM_1 convergence implies local uniform convergence at all continuity points.

Lemma 6.5.1. (local uniform convergence) *If $d_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, then (4.9) holds for each $t \notin \text{Disc}(x)$.*

Proof. For $x, t \in \text{Disc}(x)^c$ and $\epsilon > 0$ given, choose $\delta > 0$ so that $\|x(t') - x(t)\| < \epsilon$ for $|t - t'| < \delta$. Then choose $n_0 \geq 4$, $(u_n, r_n) \in \Pi_s(x_n)$ and $(u, r) \in \Pi_s(x)$ such that

$$\|u_n - u\| \vee \|r_n - r\| < (\delta \wedge \epsilon)/4$$

for all $n \geq n_0$. Let s_1, s_2, s_3 be such that $r(s_1) = t - \delta/2$, $r(s_2) = t$ and $r(s_3) = t + \delta/2$. Then $r_n(s_1) < t < \delta/4$ and $r_n(s_3) > t + \delta/4$ for all $n \geq n_0$. Hence, for all $t' \in (t - \delta/4, t + \delta/4)$ and $n \geq n_0$ there exists $s_n, s_1 < s_n < s_3$, such that $(u_n(s_n), r_n(s_n)) = (x_n(t'), t')$. Hence,

$$\begin{aligned} \|x_n(t') - x(t')\| &= \|u_n(s_n) - u(s_2)\| + \|x(t) - x(t')\| \\ &\leq \|u_n(s_n) - u(s_n)\| + \|u(s_n) - u(s_2)\| + \epsilon \\ &\leq (\delta \wedge \epsilon)/2 + 2\epsilon < 3\epsilon. \quad \blacksquare \end{aligned}$$

We next relate the modulus w_s applied to x and the modulus applied to corresponding points on the graph Γ_x . The following lemma is established in the proof of Skorohod's (1956) 2.4.1.

Lemma 6.5.2. (extending the modulus from a function to its graph) *If $(z_1, t_1), (z_2, t_2), (z_3, t_3) \in \Gamma_x$ with $0 \vee (t - \delta) \leq t_1 < t_2 < t_3 \leq (t + \delta) \wedge T$, then $\|z_2 - [z_1, z_3]\| \leq w_s(x, \delta)$.*

Proof. Suppose that $w_s(x, \delta) = \epsilon$. It suffices to show: (i) that $\|z_2 - [z_1, z_3]\| \leq \epsilon$ when $\|z'_2 - [z_1, z_3]\| \leq \epsilon$, $\|z''_2 - [z_1, z_3]\| \leq \epsilon$ and $z_2 \in [z'_2, z''_2]$ and (ii) that $\|z_2 - [z_1, z_3]\| \leq \epsilon$ when $\|z'_2 - [z'_1, z_3]\| \leq \epsilon$, $\|z_2 - [z''_1, z_3]\| \leq \epsilon$ and $z_1 \in [z'_1, z''_1]$. For (i), note that there exist $z', z'' \in [z_1, z_3]$ such that $\|z'_2 - z'\| \leq \epsilon$ and $\|z''_2 - z''\| \leq \epsilon$. Also there exists α , $0 \leq \alpha \leq 1$ such that $z_2 = \alpha z'_2 + (1 - \alpha) z''_2$. Hence $\|z_2 - (\alpha z' + (1 - \alpha) z'')\| \leq \epsilon$, which implies that

$$\|z_2 - [z', z'']\| \leq \|z_2 - [z_1, z_3]\| \leq \epsilon .$$

For (ii), note first that there exist $z' \in [z'_1, z_3]$ and $z'' \in [z''_1, z_3]$ such that $\|z_2 - z'\| \leq \epsilon$ and $\|z_2 - z''\| \leq \epsilon$. Hence, for any $z \in [z', z'']$, $\|z_2 - z\| \leq \epsilon$. The desired z lies on the intersection of $[z_1, z_3]$ and $[z', z'']$. That implies the desired conclusion. ■

Lemma 6.5.3. (asymptotic negligibility of the modulus) *For any $x \in D$, $w_s(x, \delta) \downarrow 0$ as $\delta \downarrow 0$.*

Proof. For any $\epsilon > 0$, choose $x_c \in D_c$ such that $\|x - x_c\| < \epsilon/2$, which is always possible by Theorem 6.2.2. Note that, for any $\delta > 0$,

$$w_s(x, \delta) \leq w_s(x_c, \delta) + 2\|x - x_c\| ,$$

so that

$$w_s(x, \delta) \leq w_s(x_c, \delta) + \epsilon .$$

Let η be the minimum distance between successive discontinuities in x_c . Since $w_s(x_c, \delta) = 0$ when $\delta < \eta$, $w_s(x, \delta) < \epsilon$ when $\delta < \eta$. ■

Proof of Theorem 6.5.1. Contained in the book. ■

6.5.2. WM_1 Convergence

We now establish an analog of Theorem 6.5.1 for the WM_1 topology. Several alternative characterizations of WM_1 convergence will follow directly from Theorem 6.5.1 because we will show that convergence $x_n \rightarrow x$ as $n \rightarrow \infty$ in WM_1 is equivalent to $d_p(x_n, x) \rightarrow 0$. To treat the WM_1 topology, we define another oscillation function. Let

$$w_w(x, \delta) \equiv \sup_{0 \leq t \leq T} w_w(x, t, \delta) \tag{5.7}$$

for $w_w(x, t, \delta)$ in (4.5). Recall that $w_w(x, t, \delta)$ in (4.5) is the same as $w_s(x, t, \delta)$ in (4.4) except it has the product segment $[[x(t_1), x(t_3)]]$ in (3.2) instead of the standard segment $[x(t_1), x(t_3)]$ in (3.1).

Paralleling Definition 6.3.1, let an ordered subset A of G_x of cardinality m be such that (3.9) holds, but now with the order being the order on G_x . Paralleling (3.10), let the *order-consistent distance* between A and G_x be

$$\hat{d}(A, G_x) \equiv \sup\{\|(z, t) - (z_i, t_i)\| \vee \|(z, t) - (z_{i+1}, t_{i+1})\| : (z, t) \in G_x\} \quad (5.8)$$

with the supremum being over all $(z, t) \in G_x$ such that $(z_i, t_i) \leq (z, t) \leq (z_{i+1}, t_{i+1})$ for all i , $1 \leq i \leq m-1$.

Theorem 6.5.2. (characterizations of WM_1 convergence) *The following are equivalent characterizations of $x_n \rightarrow x$ as $n \rightarrow \infty$ in (D, WM_1) :*

- (i) $d_w(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$.
- (ii) $d_p(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$.
- (iii) $x_n(t) \rightarrow x(t)$ as $n \rightarrow \infty$ for each t in a dense subset of $[0, T]$ including 0 and T , and

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w_w(x_n, \delta) = 0. \quad (5.9)$$

- (iv) $x_n(T) \rightarrow x(T)$ as $n \rightarrow \infty$; for each $t \notin \text{Disc}(x)$,

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} v(x_n, x, t, \delta) = 0 \quad (5.10)$$

for $v(x_n, x, t, \delta)$ in (4.2); and, for each $t \in \text{Disc}(x)$,

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w_w(x_n, t, \delta) = 0 \quad (5.11)$$

for $w_w(x_n, t, \delta)$ in (4.5).

- (v) for all $\epsilon > 0$ and all n sufficiently large, there exist finite ordered subsets A of G_x (in general depending on n) and A_n of G_{x_n} of common cardinality such that $\hat{d}(A, G_x) < \epsilon$, $\hat{d}(A_n, G_{x_n}) < \epsilon$ and $d^*(A, A_n) < \epsilon$ for $\hat{d}$ in (5.8) and d^* in (5.6).

Proof. (i)→(ii). Since $d_p \leq d_w$, (i)→(ii) is immediate.
(ii)↔(iii). The implication (iii)→(ii) is immediate, so we show (ii)→(iii).
By Lemma 6.5.1, $x_n^i(t) \rightarrow x^i(t)$ as $n \rightarrow \infty$ for each $t \in \text{Disc}(x^i)^c$, $1 \leq i \leq k$.
That implies that $x_n(t) \rightarrow x(t)$ as $n \rightarrow \infty$ for each $t \in \text{Disc}(x)^c$. From
Theorem 6.5.1, $d_p(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ also implies that

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w_s(x_n^i, \delta) = 0$$

for each i , $1 \leq i \leq k$, but that directly implies (5.9), because

$$\|x_n(t_2) - [[x_n(t_1), x_n(t_3)]]\| = \max_{1 \leq i \leq k} \|x_n^i(t_2) - [x_n^i(t_1), x_n^i(t_3)]\|, \quad (5.12)$$

so that

$$w_w(x_n, \delta) = \max_{1 \leq i \leq k} w_s(x_n^i, \delta). \quad (5.13)$$

(iii)↔(iv). The equivalence between (iii) and (iv) holds by the same reasoning used to establish the equivalence of (iv) and (v) in Theorem 6.5.1.

(iii)→(v). The proof of (iii)→(v) parallels the proof of (iv)→(vi) in Theorem 6.5.1, but requires some modifications. Paralleling the previous beginning, for $\epsilon > 0$ given, find $\eta < \epsilon/16$ and n_0 such that $w_w(x_n, \eta) < \epsilon/32$ for $n \geq n_0$. However, we do not next directly construct $A \in G_x$. Instead, just as with the SM_1 topology, we first construct the finite set A of Γ_x as before with the properties in the proof of Theorem 6.5.1. We denote this subset A' to distinguish it from the desired subset A of G_x . As before, for all $t_i \in S \cap A'$, let $n_1 \geq n_0$ be such that $\|x_n(t_i) - x(t_i)\| < \epsilon/32$ for all i , $1 \leq i \leq k$, and all $n \geq n_1$. We now want to construct the ordered subset A_n in G_{x_n} . For $t \in S$, the construction is as before: $(z_{n,i}, t_{n,i}) = (x_n(t_i), t_i)$. Next suppose that (??) holds. Then $(z_{n,r}, t_{n,r})$ and $(z_{n,r+j+1}, t_{n,r+j+1})$ have been defined with respect to A' . We insert points into A_n from G_{x_n} appropriately spaced in between the two points. By construction specified before (but using the product segments),

$$\begin{aligned} & \|[(x_n(t_r), t_r), (x_n(t_{r+j+1}), t_{r+j+1})] \\ & - [(x(t_r), t_r), (x(t_{r+j+1}), t_{r+j+1})]\| < \epsilon/32 \end{aligned} \quad (5.14)$$

and

$$\|[(x(t_r), t_r), (x(t_{r+j+1}), t_{r+j+1})] - [(x(t-), t), (x(t), t)]\| < \epsilon/32. \quad (5.15)$$

To simplify the discussion, suppose that $x^i(t-) \leq x^i(t)$ for all i . (This is without loss of generality after redefining the order.) Consider an arbitrary nondecreasing (in the order on G_{x_n}) continuous curve in G_{x_n} from

$(z_{n,r}, t_{n,r})$ to $(z_{n,r+j+1}, t_{n,r+j+1})$. Let $(z'_{n,r+1}, t'_{n,r+1})$ be the first point on this curve for which the i^{th} coordinate first reaches $z_{n,r}^i + \epsilon/4$ for some i . Given $(z_{n,r+k}, t_{n,r+k})$, let $(z_{n,r+k+1}, t_{n,r+k+1})$ be the next point on the curve at which the i^{th} coordinate first reaches $z_{n,r+k}^i + \epsilon/4$ for some i . Since $x^i(t-) \leq x^i(t)$ for all i and since $w_w(x_n, \eta) < \epsilon/32$, no coordinate of the curve in G_{x_n} can decrease by more than $\epsilon/32$ over any subinterval, and thus from one point to the next in A_n . Continue in this manner for at most finitely many steps until the end point $(z_{n,r+j+1}, t_{n,r+j+1})$ is reached. The distance between successive points is $\epsilon/4$, while the distance between the last point inserted and $(z_{n,r+j+1}, t_{n,r+j+1})$ is less than $\epsilon/4$. Delete the first and last point inserted, so that all distances between successive points are between $\epsilon/4$ and $\epsilon/2$. In general, the number of inserted points is some finite number, not necessarily equal to j . These points are ordered, since they lie on the non-decreasing continuous curve through G_{x_n} . For each $t \in \text{Disc}(x, \epsilon/2)$, let A_n contain these specified points. This construction yields $\hat{d}(A_n, G_{x_n}) < \epsilon/2$. For $t \notin \text{Disc}(x, \epsilon/2)$, let A contain the points already constructed in A' . It remains to construct the points in A for $t \in \text{Disc}(x, \epsilon/2)$. For this purpose, we use the points in A_n associated with t . Again, to simplify the discussion, suppose that $x^i(t-) \leq x^i(t)$ for all i . With this ordering, we let

$$z_{r+k}^i = x^i(t-) \vee \max_{1 \leq l \leq k} z_{n,r+l}^i \wedge x^i(t)$$

for each k and i . This definition guarantees that the points (z_{r+k}, t) belong to G_x and are ordered. Moreover, $\hat{d}(A, G_x) < \epsilon$. Finally, we must have $d^*(A, A_n) < \epsilon$, because otherwise the condition $w_w(x_n, \eta) < \epsilon/32$ would be violated.

(v) $\rightarrow$ (i). Suppose that the conditions in (v) hold and let $\epsilon > 0$ be given. Construct the finite subsets A and A_n with the specified properties. Let (u, r) and (u_n, r_n) be arbitrary parametric representations of G_x and G_{x_n} such that there are points s_i in $S \subseteq [0, 1]$ such that both $(u(s_i), r(s_i)) = (z_i, t_i) \in A$ and $(u_n(s_i), r_n(s_i)) = (z_{n,i}, t_{n,i}) \in A_n$. Since A and A_n are ordered subsets of G_x and G_{x_n} , respectively that construction is possible. Finally, for any s , $0 < s < 1$, there is $s_i \in S$ such that $s_i \leq s < s_{i+1}$ and

$$\begin{aligned} & \|u_n(s) - u(s)\| \vee \|r_n(s) - r(s)\| \leq \|(u_n(s), r_n(s)) - (u_n(s_i), r_n(s_i))\| \\ & \quad + \|(u_n(s_i), r_n(s_i)) - (u(s_i), r(s_i))\| + \|(u(s_i), r(s_i)) - (u(s), r(s))\| \\ & \leq \hat{d}(A_n, G_{x_n}) + d^*(A, A_n) + \hat{d}(A, G_x) \leq 3\epsilon. \quad \blacksquare \end{aligned}$$

6.6. Strengthening the Mode of Convergence

Section 12.6 of the book applies the characterizations of M_1 convergence in previous sections to establish conditions under which the mode of convergence can be strengthened: We find conditions under which WM_1 convergence can be replaced by SM_1 convergence. Most of the material appears in the book.

We use the following Lemma.

Lemma 6.6.1. (modulus bound for (x_n, y_n)) For $x_n \in D([0, T], \mathbb{R}^k)$, $y_n, y \in D([0, T], \mathbb{R}^l)$, $t \in [0, T]$ and $\delta > 0$,

$$w_s((x_n, y_n), t, \delta) \leq w_s(x_n, t, \delta) + 2v(y_n, y, t, \delta).$$

Proof. For $(t - \delta) \vee 0 \leq t_1 < t_2 < t_3 \leq (t + \delta) \wedge T$,

$$\begin{aligned} \|(x_n, y_n)(t_2) - [(x_n, y_n)(t_1), (x_n, y_n)(t_3)]\| & \\ & \leq \|(x_n, y_n)(t_2) - [(x_n(t_1), y(t_1)), (x_n(t_3), y(t_3))]\| \\ & \quad + (\|y_n(t_1) - y(t_1)\| \vee \|y_n(t_3) - y(t_3)\|) \\ & \leq \|x_n(t_2) - [x_n(t_1), x_n(t_3)]\| \vee \|y_n(t_2) - y(t_2)\| \\ & \quad + (\|y_n(t_1) - y(t_1)\| \vee \|y_n(t_3) - y(t_3)\|) \\ & \leq \|x_n(t_2) - [x_n(t_1), x_n(t_3)]\| + 2v(y_n, y, t, \delta). \quad \blacksquare \end{aligned}$$

Theorem 6.6.1. (extending SM_1 convergence to product spaces) Suppose that $d_s(x_n, x) \rightarrow 0$ in $D([0, T], \mathbb{R}^k)$ and $d_s(y_n, y) \rightarrow 0$ in $D([0, T], \mathbb{R}^l)$ as $n \rightarrow \infty$. If

$$Disc(x) \cap Disc(y) = \emptyset.$$

then

$$d_s((x_n, y_n), (x, y)) \rightarrow 0 \text{ in } D([0, T], \mathbb{R}^{k+l}) \text{ as } n \rightarrow \infty.$$

The proof is in the book.

6.7. Characterizing Convergence with Mappings

In this section we focus on alternative characterizations of SM_1 convergence using mappings.

6.7.1. Linear Functions of the Coordinates

The strong topology SM_1 differs from the weak topology WM_1 by the behavior of linear functions of the coordinates. Example ?? shows that linear functions of the coordinates are not continuous in the product topology (there $(x_n^1 - x_n^2) \not\rightarrow (x^1 - x^2)$ as $n \rightarrow \infty$), but they are in the strong topology, as we now show. Note that there is no subscript on d on the left in (7.1) below because ηx is real valued.

Theorem 6.7.1. (Lipschitz property of linear functions of the coordinate functions) *For any $x_1, x_2 \in D([0, T], \mathbb{R}^k)$ and $\eta \in \mathbb{R}^k$,*

$$d(\eta x_1, \eta x_2) \leq (\|\eta\| \vee 1) d_s(x_1, x_2) . \quad (7.1)$$

Proof. Pick an arbitrary $\epsilon > 0$ and choose $(u_j, r_j) \in \Pi_s(x_j)$ for $j = 1, 2$ such that

$$\|u_1 - u_2\| \vee \|r_1 - r_2\| < d_s(x_1, x_2) + \epsilon ,$$

which is possible by the definition (3.7). Because $\eta u_j \in \Pi(\eta x_j)$ for $j = 1, 2$, by Lemma 6.3.4,

$$\begin{aligned} d(\eta x_1, \eta x_2) &\leq \|\eta u_1 - \eta u_2\| \vee \|r_1 - r_2\| \\ &\leq \|r_1 - r_2\| \vee \|u_1 - u_2\| \|\eta\| \\ &\leq (\|\eta\| \vee 1) (d_s(x_1, x_2) + \epsilon) . \end{aligned}$$

Since ϵ was arbitrary, (7.1) is established. ■

We now obtain a sufficient condition for addition to be continuous on $(D, d_s) \times (D, d_s)$, which is analogous to the J_1 result in Theorem 4.1 of Whitt (1980).

Corollary 6.7.1. (SM_1 -continuity of addition) *If $d_s(x_n, x) \rightarrow 0$ and $d_s(y_n, y) \rightarrow 0$ in $D([0, T], \mathbb{R}^k)$ and*

$$Disc(x) \cap Disc(y) = \phi ,$$

then

$$d_s(x_n + y_n, x + y) \rightarrow 0 \text{ in } D([0, T], \mathbb{R}^k) .$$

Proof. First apply Theorem 6.6.1 to get $d_s((x_n, y_n), (x, y)) \rightarrow 0$ in $D([0, T], \mathbb{R}^{2k})$. Then apply Theorem 6.7.1. ■

Remark 6.7.1. *Measurability of addition.* The measurability of addition on $(D, d_s) \times (D, d_s)$ holds because the Borel σ -field coincides with the Kolmogorov σ -field. It also follows from part of the proof of Theorem 4.1 of Whitt (1980). ■

In Theorem 6.7.1 we showed that linear functions of the coordinates are Lipschitz in the SM_1 metric. We now apply Theorem 6.5.1 to show that convergence in the SM_1 topology is characterized by convergence of all such linear functions of the coordinates.

Theorem 6.7.2. (characterization of SM_1 convergence by convergence of all linear functions) *There is convergence $x_n \rightarrow x$ in $D([0, T], \mathbb{R}^k)$ as $n \rightarrow \infty$ in the SM_1 topology if and only if $\eta x_n \rightarrow \eta x$ in $D([0, T], \mathbb{R}^1)$ as $n \rightarrow \infty$ in the M_1 topology for all $\eta \in \mathbb{R}^k$.*

Proof. One direction is covered by Theorem 6.7.1. Suppose that $x_n \not\rightarrow x$ as $n \rightarrow \infty$ in SM_1 . Then apply part (v) of Theorem 6.5.1 to deduce that $\eta x_n \not\rightarrow \eta x$ as $n \rightarrow \infty$ for some η . Note that $\|a\| > 0$ for $a \in \mathbb{R}^k$ if and only if $|a_i| > 0$ for some i . Also, $\|a - A\| > 0$ for $A \subseteq \mathbb{R}^k$ if and only if $|a_i - A_i| > 0$ for some i in $\mathbb{R}$ for some $\eta \in \mathbb{R}^k$, where $\eta A = \{\eta b : b \in A\}$. ■

We can get convergence of sums under more general conditions than in Corollary 6.7.1. It suffices to have the jumps of x^i and y^i have common sign for all i . We can express this property by the condition

$$(x^i(t) - x^i(t-))(y^i(t) - y^i(t-)) \geq 0 \quad (7.2)$$

for all t , $0 \leq t \leq T$, and all i , $1 \leq i \leq k$.

Theorem 6.7.3. (continuity of addition at limits with jumps of common sign) *If $x_n \rightarrow x$ and $y_n \rightarrow y$ in $D([0, T], \mathbb{R}^k, SM_1)$ and if condition (7.2) above holds, then*

$$x_n + y_n \rightarrow x + y \quad \text{in} \quad D([0, T], \mathbb{R}^k, SM_1) .$$

Proof. The proof is in the book.

6.7.2. Visits to Strips

In Sections (2.2.7)–(2.2.13) of Skorohod (1956), convenient characterizations of convergence in each topology are given for real-valued functions. We can apply Theorem 6.7.2 to develop associated characterizations for $\mathbb{R}^k$ -valued functions. For each $x \in D([0, T], \mathbb{R}^1)$, $0 \leq t_1 < t_2 \leq T$ and, for each $a < b$ in $\mathbb{R}$, let $v_{t_1, t_2}^{a, b}(x)$ be the number of visits to the strip $[a, b]$ on the interval $[t_1, t_2]$; i.e., $v_{t_1, t_2}^{a, b}(x) = k$ if it is possible to find k (but not $k + 1$) points t'_i such that $t_1 < t'_1 < \cdots < t'_k \leq t_2$ such that either

$$x(t_1) \in [a, b], \quad x(t'_1) \notin [a, b], \quad x(t'_2) \in [a, b], \dots,$$

or

$$x(t_1) \notin [a, b], \quad x(t'_1) \in [a, b], \quad x(t'_2) \notin [a, b], \dots$$

We say that $x \in D([0, T], \mathbb{R})$ has a *local maximum (minimum) value at t relative to (t_1, t_2)* in $(0, T)$ if $t_1 < t < t_2$ and either

$$(i) \quad \sup\{x(s) : t_1 \leq s \leq t_2\} \leq x(t) \quad (\inf\{x(s) : t_1 \leq s \leq t_2\} \geq x(t))$$

or

$$(ii) \quad \sup\{x(s) : t_1 \leq s \leq t_2\} \leq x(t-) \quad (\inf\{x(s) : t_1 \leq s \leq t_2\} \geq x(t-)) .$$

We say that x has a *local maximum (minimum) value at t* if it has a local maximum (minimum) value at t relative to some interval (t_1, t_2) with $t_1 < t < t_2$. We call local maximum and minimum values *local extreme values*.

Lemma 6.7.1. (local extreme values) *Any $x \in D([0, T], \mathbb{R})$ has at most countably many local extreme values.*

Proof. For each n , let $\{t_{n,i}\}$ be a finite collection of points in $[0, T]$, including 0 and T . Let $\{t_{n,i}\}$ be a subcollection of $\{t_{n+1,i}\}$ for each n and let the minimum distance between points in $\{t_{n,i}\}$ be ϵ_n , where $\epsilon_n \downarrow 0$ as $n \rightarrow \infty$. Note that there is one local maximum value and one local minimum value of x relative to the interval endpoints in each interval $[t_{n,i}, t_{n,i+1})$, where $t_{n,i}$ and $t_{n,i+1}$ are successive points in $\{t_{n,i}\}$. Hence the total number of extreme values of x relative to $\{t_{n,i}\}$ is countably infinite. Next note that any extreme value of x is contained in this set. To see this, suppose that b is an extreme value of x at t relative to the interval (t_1, t_2) . Then, for sufficiently large n , there is an interval $(t_{n,i}, t_{n,i+1})$ such that $t_1 \leq t_{n,i} < t < t_{n,i+1} \leq t_2$, so that b is an extreme value of x within $(t_{n,i}, t_{n,i+1})$. ■

If b is not a local extreme value of x , then x crosses level b whenever x hits b ; i.e., if b is not a local extreme value and if $x(t) = b$ or $x(t-) = b$, then for every t_1, t_2 with $t_1 < t < t_2$ there exist t'_1, t'_2 with $t_1 < t'_1, t'_2 < t_2$ such that $x(t'_1) < b$ and $x(t'_2) > b$. This property implies the following lemma.

Lemma 6.7.2. *Consider an interval $[t_1, t_2]$ with $0 < t_1 < t_2 < T$. If $x(t_i) \notin \{a, b\}$ for $i = 1, 2$ and a, b are not local extreme values of x , then x crosses one of the levels a and b at each of the $v_{t_1, t_2}^{a, b}(x)$ visits to the strip $[a, b]$ in $[t_1, t_2]$.*

Theorem 6.7.4. (characterization of SM_1 convergence in terms of convergence of number of visits to strips) *There is convergence $d_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ in $D([0, T], \mathbb{R}^k)$ if and only if*

$$v_{t_1, t_2}^{a, b}(\eta x_n) \rightarrow v_{t_1, t_2}^{a, b}(\eta x) \quad \text{as } n \rightarrow \infty$$

for all $\eta \in \mathbb{R}^k$, all points $t_1, t_2 \in \{T\} \cup \text{Disc}(x)^c$ with $t_1 < t_2$ and almost all a, b with respect to Lebesgue measure.

Proof. By Theorem 6.7.2, it suffices to establish the result for $\mathbb{R}$ -valued functions. First, suppose that $x_n \rightarrow x$ as $n \rightarrow \infty$ in $D([0, T], \mathbb{R}, M_1)$. Suppose that a and b are not local extreme values of x . Let $t_1, t_2 \in \text{Disc}(x)^c$ and suppose that $x(t_1), x(t_2) \notin \{a, b\}$. Then, for sufficiently large n , by Lemma 6.7.2, $v_{t_1, t_2}^{a, b}(x_n) = v_{t_1, t_2}^{a, b}(x)$. Since there are at most countably many “bad” a, b for any x , $v_{t_1, t_2}^{a, b}(x_n) \rightarrow v_{t_1, t_2}^{a, b}(x)$ for almost all a, b with respect to Lebesgue measure. On the other hand, suppose that $v_{t_1, t_2}^{a, b}(x_n) \rightarrow v_{t_1, t_2}^{a, b}(x)$ for all $t_1, t_2 \in \text{Disc}(x)^c$ and for almost all a, b . We will show that characterization (v) of SM_1 convergence in Theorem 6.5.1 holds. For x, t and $\epsilon > 0$ given, find η such that $v(x, [t - \eta, t]) < \epsilon/2$ and $v(x, [t, t + \eta]) < \epsilon/2$. First suppose that $t \in \text{Disc}(x)^c$. Then $v_{t_1, t_2}^{a, b}(x) = 0$ for $t_1, t_2 \in \text{Disc}(x)^c$, $t - \eta < t_1 < t < t_2 < t + \eta$ and all (a, b) with $a < x(t) - \epsilon/2 < x(t) + \epsilon/2 < b$. By assumption, for all suitably large n , $v_{t_1, t_2}^{a', b'}(x_n) = 0$ for some a', b' with

$$x(t) - \epsilon < a' < x(t) - \epsilon/2 < x(t) + \epsilon/2 < b' < x(t) + \epsilon.$$

By the argument above, we can show that, for a time interval before t , x_n and x are first in a neighborhood of $x(t-)$ and then leave. Afterwards, x_n and x enter the neighborhood of $x(t)$ and stay there for a short interval after t . To see this, let t_1 and t_2 be as above and then find a_1, b_1, a_2, b_2 such that

$$\begin{aligned} x(t-) - \epsilon < a_1 < x(t-) - \epsilon/2, \quad x(t-) + \epsilon/2 < b_1 < x(t) + \epsilon \\ x(t) - \epsilon < a_2 < x(t) - \epsilon/2, \quad x(t) + \epsilon/2 < b_2 < x(t) + \epsilon, \end{aligned}$$

$v_{t_1, t_2}^{a_1, b_1}(x_n) \rightarrow v_{t_1, t_2}^{a_1, b_1}(x) = 1$ and $v_{t_1, t_2}^{a_2, b_2}(x_n) \rightarrow v_{t_1, t_2}^{a_2, b_2}(x) = 1$. that implies that $v(x_n, x, t, \delta) < \epsilon$ for $\delta < \min\{|t - t_1|, |t - t_2|\}$. Next suppose that $t \in \text{Disc}(x)$. Let t_1, t_2 be as above. Find a_1, b_1, a_2, b_2 such that

$$x(t-) - \epsilon < a_1 < x(t-) - \epsilon/2 < x(t-) + \epsilon/2 < b_1 < x(t-) + \epsilon,$$

$$x(t) - \epsilon < a_2 < x(t) - \epsilon/2 < x(t) + \epsilon/2 < b < x(t) + \epsilon,$$

$v_{t_1, t_2}^{a_1, b_1}(x_n) \rightarrow v_{t_1, t_2}^{a_1, b_1}(x) = 1$ and $v_{t_1, t_2}^{a_2, b_2}(x_n) \rightarrow v_{t_1, t_2}^{a_2, b_2}(x) = 1$. It remains to show that x_n cannot fluctuate significantly between $x(t-)$ and $x(t)$. To be definite, suppose that $x(t-) < x(t)$ and suppose that $\epsilon < x(t) - x(t-)$. Then for almost all a, b with

$$x(t-) + \epsilon/2 < a < b < x(t) - \epsilon/2,$$

$$v_{t_1, t_2}^{a, b}(x_n) \rightarrow v_{t_1, t_2}^{a, b}(x) = 2 \quad \text{as } n \rightarrow \infty.$$

That implies that $w_s(x_n, x, t, \delta) \rightarrow 0$ as $n \rightarrow \infty$ for $\delta < \min\{|t_1 - t|, |t - t_2|\}$, which completes the proof. ■

6.8. Topological Completeness

In this section we exhibit a complete metric topologically equivalent to the incomplete metric d_s in (3.7) inducing the SM_1 topology. Since a product metric defined as in (3.13) inherits the completeness of the component metrics, we also succeed in constructing complete metrics inducing the associated product topology. We make no use of the complete metrics beyond showing that the topology is topologically complete. Another approach to topological completeness would be to show that D is homeomorphic to a G_δ subset of a complete metric space, as noted in Section 11.2 of the book.

In our construction of complete metrics, we follow the argument used by Prohorov (1956, Appendix 1) to show that the J_1 topology is topologically complete; we incorporate an oscillation function into the metric. For M_1 , we use $w_s(x, \delta)$ in (5.1). Since $w_s(x, \delta) \rightarrow 0$ as $\delta \rightarrow 0$ for each $x \in D$, we need to appropriately “inflate” differences for small δ . For this purpose, let

$$\hat{w}_s(x, z) \equiv \begin{cases} w_s(x, e^z), & z < 0 \\ w_s(x, 1), & z \geq 1. \end{cases} \quad (8.1)$$

Since $w_s(x, \delta)$ is nondecreasing in δ , $\hat{w}_s(x, z)$ is nondecreasing in z . Note that $\hat{w}_s(x, z)$ as a function of z has the form of a cumulative distribution

function (cdf) of a finite measure. On such cdf's, the Lévy metric λ is known to be a complete metric inducing the topology of pointwise convergence at all continuity points of the limit; i.e.,

$$\lambda(F_1, F_2) \equiv \inf\{\epsilon > 0 : F_2(x - \epsilon) - \epsilon \leq F_1(x) \leq F_2(x + \epsilon) + \epsilon\} . \quad (8.2)$$

The Helly selection theorem, p. 267 of Feller (1971), can be used to show that the metric λ is complete.

Thus, our new metric is

$$\hat{d}_s(x_1, x_2) \equiv d_s(x_1, x_2) + \lambda(\hat{w}_s(x_1, \cdot), \hat{w}_s(x_2, \cdot)) . \quad (8.3)$$

Theorem 6.8.1. (a complete SM_1 metric) *The metric $\hat{d}_s$ on D in (8.3) is complete and topologically equivalent to d_s .*

Proof. To show topological equivalence of $\hat{d}_s$ and d_s , it suffices to show that $\lambda(\hat{w}_s(x_n, \cdot), \hat{w}_s(x, \cdot)) \rightarrow 0$ as $n \rightarrow \infty$ whenever $d_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$. However, if $d_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, then $w_s(x_n, \delta) \rightarrow w_s(x, \delta)$ as $n \rightarrow \infty$ at all δ which are continuity points of $w_s(x, \delta)$. (See Lemma 6.8.1 below.) That in turn implies that $\hat{w}_s(x_n, z) \rightarrow \hat{w}_s(x, z)$ as $n \rightarrow \infty$ for all z which are continuity points of $\hat{w}_s(x, z)$. However, such convergence is equivalent to convergence under λ . Next, suppose that a sequence $\{x_n\}$ is fundamental under $\hat{d}_s$, i.e., $\hat{d}_s(x_m, x_n) \rightarrow 0$ as $m, n \rightarrow \infty$. It follows that $\{x_n(t) : 0 \leq t \leq T, n \geq 1\}$ is compact. Hence, there exists a countable dense set N of $[0, T]$, including 0 and T , and a subsequence $\{x_{n_k}\}$ such that $x_{n_k}(t) \rightarrow x(t)$ as $n_k \rightarrow \infty$ for all $t \in N$, where x is some $\mathbb{R}^k$ -valued function on $[0, T]$. At the same time, since λ is known to be a complete metric, there must exist a distribution function F such that

$$\lim_{n \rightarrow \infty} \lambda(\hat{w}_s(x_n, \cdot), F) = 0 ,$$

which implies that

$$\lim_{\delta \rightarrow 0} \overline{\lim}_{n \rightarrow \infty} w_s(x_n, \delta) = 0 .$$

However, Theorem ?? and Corollary ?? imply that there exists $\bar{x} \in D$ (with $\bar{x}$ not necessarily x) such that $d_s(x_{n_k}, \bar{x}) \rightarrow 0$ as $n_k \rightarrow \infty$. Since $d_s(x_n, \bar{x}) \leq d_s(x_n, x_{n_k}) + d_s(x_{n_k}, \bar{x})$ and $d_s(x_m, x_n) \rightarrow 0$ as $m, n \rightarrow \infty$, $d_s(x_n, \bar{x}) \rightarrow 0$ as $n \rightarrow \infty$. ■

To complete the proof of Theorem 6.8.1, we need the following lemma.

Lemma 6.8.1. (continuity of SM_1 modulus) *If $d_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, then $w_s(x_n, \delta) \rightarrow w_s(x, \delta)$ as $n \rightarrow \infty$ for each δ that is a continuity point of $w_s(x, \delta)$.*

Proof. Let δ be a continuity point of $w_s(x, \delta)$. Then, for each $\epsilon_1 > 0$, there is $\epsilon_2 > 0$ such that $w_s(x, \delta - \epsilon_2) \geq w_s(x, \delta) - \epsilon_1$. For δ , ϵ_1 and ϵ_2 given, it is possible to choose continuity points t , t_1 , t_2 and t_3 of x such that

$$(t - \delta) \vee 0 \leq t_1 \leq t_2 \leq t_3 \leq (t + \delta) \wedge T \quad (8.4)$$

and

$$\|x(t_2) - [x(t_1), x(t_3)]\| \geq w_s(x, \delta - \epsilon_2) - \epsilon_1 \geq w_s(x, \delta) - 2\epsilon_1 .$$

Since $d_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, $x_n(t_j) \rightarrow x(t_j)$ as $n \rightarrow \infty$ for $j = 1, 2, 3$. Hence, there exists n_0 such that, for all $n \geq n_0$,

$$\|x_n(t_2) - [x_n(t_1), x_n(t_3)]\| \geq w_s(x, \delta) - 3\epsilon_2 .$$

However,

$$w_s(x_n, \delta) \geq \|x_n(t_2) - [x_n(t_1), x_n(t_3)]\| ,$$

so that $w_s(x_n, \delta) \geq w_s(x, \delta) - 3\epsilon_2$. Since ϵ_2 can be made arbitrarily small,

$$\lim_{n \rightarrow \infty} w_s(x_n, \delta) \geq w_s(x, \delta) . \quad (8.5)$$

We now establish an inequality in the other direction. Since δ is a continuity point of $w_s(x, \delta)$, for any $\epsilon_1 > 0$ there exists $\epsilon_2 > 0$ so that $w_s(x, \delta + \epsilon_2) \leq w_s(x, \delta) + \epsilon_1$. We can choose t_n , t_{n1} , t_{n2} and t_{n3} so that

$$(t_n - \delta) \vee 0 \leq t_{n1} \leq t_{n2} \leq t_{n3} \leq (t_n + \delta) \wedge T$$

and

$$\|x_n(t_{n2}) - [x_n(t_{n1}), x_n(t_{n3})]\| \geq w_s(x_n, \delta) - \epsilon_2$$

for all n . There thus exists a subsequence $\{n_k\}$ such that $t_{n_k} \rightarrow t$ and $t_{n_k j} \rightarrow t_j$, $j = 1, 2, 3$, (8.4) holds and $\|x_{n_k}(t_{n_k j}) - z_j\| \rightarrow 0$ as $n_k \rightarrow \infty$. Moreover, since x and x_n , $n \geq 1$, are right-continuous for all n , we can have t_1 , t_2 and t_3 be continuity points of x with

$$(t - (\delta + \epsilon_2)) \vee 0 \leq t_1 \leq t_2 \leq t_3 \leq (t + (\delta + \epsilon_2)) \wedge T .$$

Then $\|x_{n_k}(t_{n_k j}) - x(t_j)\| \rightarrow 0$ as $n_k \rightarrow \infty$. Hence, there is n_0 such that, for all $n_k \geq n_0$,

$$\begin{aligned} \|x(t_2) - [x(t_1), x(t_3)]\| &\geq \|x_{n_k}(t_{n_k 2}) - [x_{n_k}(t_{n_k 1}), x_{n_k}(t_{n_k 3})]\| - \epsilon_2 \\ &\geq w_s(x_n, \delta) - 2\epsilon_2 . \end{aligned} \quad (8.6)$$

However,

$$w_s(x, \delta) + \epsilon_1 \geq w_s(x, \delta + \epsilon_2) \geq \|x(t_2) - [x(t_1), x(t_3)]\| . \quad (8.7)$$

Combining (8.6) and (8.7), we obtain

$$w_s(x, \delta) \geq w_s(x_n, \delta) - \epsilon_1 - 2\epsilon_2 .$$

Since ϵ_1 and ϵ_2 can be made arbitrarily small,

$$\overline{\lim}_{n \rightarrow \infty} w_s(x_n, \delta) \leq w_s(x, \delta) . \quad (8.8)$$

Combining (8.5) and (8.8) completes the proof. ■

6.9. Non-Compact Domains

It is often convenient to consider the function space $D([0, \infty), \mathbb{R}^k)$ with domain $[0, \infty)$ instead of $[0, T]$. More generally, we may consider the function space $D(I, \mathbb{R}^k)$, where I is a subinterval of the real line. Common cases besides $[0, \infty)$ are $(0, \infty)$ and $(-\infty, \infty) \equiv \mathbb{R}$.

Given the function space $D(I, \mathbb{R}^k)$ for any subinterval I , we define convergence $x_n \rightarrow x$ with some topology to be convergence in $D([a, b], \mathbb{R}^k)$ with that same topology for the restrictions of x_n and x to the compact interval $[a, b]$ for all points a and b that are elements of I and either boundary points of I or are continuity points of the limit function x . For example, for $I = [c, d)$ with $-\infty < c < d < \infty$, we include $a = c$ but exclude $b = d$; for $I = [c, d]$, we include both c and d .

For simplicity, we henceforth consider only the special case in which $I = [0, \infty)$. In that setting, we can equivalently define convergence $x_n \rightarrow x$ as $n \rightarrow \infty$ in $D([0, \infty), \mathbb{R}^k)$ with some topology to be convergence $x_n \rightarrow x$ as $n \rightarrow \infty$ in $D([0, t], \mathbb{R}^k)$ with that topology for the restrictions of x_n and x to $[0, t]$ for $t = t_k$ for each t_k in some sequence $\{t_k\}$ with $t_k \rightarrow \infty$ as $k \rightarrow \infty$, where $\{t_k\}$ can depend on x . It suffices to let t_k be continuity points of the limit function x ; for the J_1 topology, see Lindvall (1973),

Whitt (1980) and Jacod and Shiryaev (1987). We will discuss only the SM_1 topology here, but the discussion applies to the other non-uniform topologies as well. We also will omit most proofs.

As a first step, we consider the case of closed bounded intervals $[t_1, t_2]$. The space $D([t_1, t_2], \mathbb{R}^k)$ is essentially the same as (homeomorphic to) the space $D([0, T], \mathbb{R}^k)$ already studied, but we want to look at the behavior

as we change the interval $[t_1, t_2]$. For $[t_3, t_4] \subseteq [t_1, t_2]$, we consider the restriction of x in $D([t_1, t_2], \mathbb{R}^k)$ to $[t_3, t_4]$, defined by

$$r_{t_3, t_4} : D([t_1, t_2], \mathbb{R}^k) \rightarrow D([t_3, t_4], \mathbb{R}^k)$$

with $r_{t_3, t_4}(x)(t) = x(t)$ for $t_3 \leq t \leq t_4$. Let d_{t_1, t_2} be the metric d_s on $D([t_1, t_2], \mathbb{R}^k)$. We want to relate the distance $d_{t_1, t_2}(x_1, x_2)$ and convergence $d_{t_1, t_2}(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ for different domains. We first state a result enabling us to go from the domains $[t_1, t_2]$ and $[t_2, t_3]$ to $[t_1, t_3]$ when $t_1 < t_2 < t_3$.

Lemma 6.9.1. (metric bounds) *For $0 \leq t_1 < t_2 < t_3$ and $x_1, x_2 \in D([t_1, t_3], \mathbb{R}^k)$,*

$$d_{t_1, t_3}(x_1, x_2) \leq d_{t_1, t_2}(x_1, x_2) \vee d_{t_2, t_3}(x_1, x_2) .$$

We now observe that there is an equivalence of convergence provided that the internal boundary point is a continuity point of the limit function.

Lemma 6.9.2. *For $0 \leq t_1 < t_2 < t_3$ and $x, x_n \in D([t_1, t_3], \mathbb{R}^k)$, with $t_2 \in \text{Disc}(x)^c$, $d_{t_1, t_3}(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ if and only if $d_{t_1, t_2}(x_n, x) \rightarrow 0$ and $d_{t_2, t_3}(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$.*

For $x \in D([0, T], \mathbb{R}^k)$ and $0 \leq t_1 < t_2 \leq T$, let $r_{t_1, t_2} : D([0, T], \mathbb{R}^k) \rightarrow D([t_1, t_2], \mathbb{R}^k)$ be the restriction map, defined by $r_{t_1, t_2}(x)(s) = x(s)$, $t_1 \leq s \leq t_2$.

Corollary 6.9.1. (continuity of restriction maps) *If $x_n \rightarrow x$ as $n \rightarrow \infty$ in $D([0, T], \mathbb{R}^k, SM_1)$ and if $t_1, t_2 \in \text{Disc}(x)^c$, then*

$$r_{t_1, t_2}(x_n) \rightarrow r_{t_1, t_2}(x) \text{ as } n \rightarrow \infty \text{ in } D([t_1, t_2], \mathbb{R}^k, SM_1) .$$

Let $r_t : D([0, \infty), \mathbb{R}^k) \rightarrow D([0, t], \mathbb{R}^k)$ be the restriction map with $r_t(x)(s) = x(s)$, $0 \leq s \leq t$. Suppose that $f : D([0, \infty), \mathbb{R}^k) \rightarrow D([0, \infty), \mathbb{R}^k)$ and $f_t : D([0, t], \mathbb{R}^k) \rightarrow D([0, t], \mathbb{R}^k)$ for $t > 0$ are functions with

$$f_t(r_t(x)) = r_t(f(x))$$

for all $x \in D([0, \infty), \mathbb{R}^k)$ and all $t > 0$. We then call the functions f_t restrictions of the function f .

Theorem 6.9.1. (continuity from continuous restrictions) *Suppose that $f : D([0, \infty), \mathbb{R}^k) \rightarrow D([0, \infty), \mathbb{R}^l)$ has continuous restrictions f_t with some topology for all $t > 0$. Then f itself is continuous in that topology.*

Proof. Suppose that $x_n \rightarrow x$ as $n \rightarrow \infty$ in the specified topology. That means that $r_{t_m}(x_n) \rightarrow r_{t_m}(x)$ as $n \rightarrow \infty$ for some sequence $\{t_m\}$ with $t_m \rightarrow \infty$, possibly depending on x . Since f has continuous restrictions,

$$r_{t_m}(f(x_n)) = f_{t_m}(r_{t_m}(x_n)) \rightarrow f_{t_m}(r_{t_m}(x)) = r_{t_m}(f(x))$$

as $n \rightarrow \infty$ for all m , which implies that $f(x_n) \rightarrow f(x)$ as $n \rightarrow \infty$ in the specified topology. ■

No more material has been deleted from Section 12.9 of the book.

6.10. Strong and Weak M_2 Topologies

We now define strong and weak versions of Skorohod's M_2 topology. In Section 6.11 we will show that it is possible to define the M_2 topologies by a minor modification of the definitions in Section 6.3, in particular, by simply using parametric representations in which only r is nondecreasing instead of (u, r) , but now we will use Skorohod's (1956) original approach, and relate it to the Hausdorff metric on the space of graphs.

The weak topology will be defined just like the strong, except it will use the thick graphs G_x instead of the thin graphs Γ_x . In particular, let

$$\mu_s(x_1, x_2) \equiv \sup_{(z_1, t_1) \in \Gamma_{x_1}} \inf_{(z_2, t_2) \in \Gamma_{x_2}} \{\|(z_1, t_1) - (z_2, t_2)\|\} \quad (10.1)$$

and

$$\mu_w(x_1, x_2) \equiv \sup_{(z_1, t_1) \in G_{x_1}} \inf_{(z_2, t_2) \in G_{x_2}} \{\|(z_1, t_1) - (z_2, t_2)\|\} . \quad (10.2)$$

Following Skorohod (1956), we say that $x_n \rightarrow x$ as $n \rightarrow \infty$ for a sequence or net $\{x_n\}$ in the strong M_2 topology, denoted by SM_2 if $\mu_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$. Paralleling that, we say that $x_n \rightarrow x$ as $n \rightarrow \infty$ in the weak M_2 topology, denoted by WM_2 , if $\mu_w(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$. We say that $x_n \rightarrow x$ as $n \rightarrow \infty$ in the product topology if $\mu_s(x_n^i, x^i) \rightarrow 0$ (or equivalently $\mu_w(x_n^i, x^i) \rightarrow 0$) as $n \rightarrow \infty$ for each i , $1 \leq i \leq k$.

We can also generate the SM_2 and WM_2 topologies using the Hausdorff metric in equation 5.2 of Section 11.5 in the book. As in equation (5.4) in Section 11.5 of the book, for $x_1, x_2 \in D$,

$$m_s(x_1, x_2) \equiv m_H(\Gamma_{x_1}, \Gamma_{x_2}) = \mu_s(x_1, x_2) \vee \mu_s(x_2, x_1) , \quad (10.3)$$

$$m_w(x_1, x_2) \equiv m_H(G_{x_1}, G_{x_2}) = \mu_w(x_1, x_2) \vee \mu_w(x_2, x_1) \quad (10.4)$$

and

$$m_p(x_1, x_2) \equiv \max_{1 \leq i \leq k} m_s(x_1^i, x_2^i) . \quad (10.5)$$

We will show that the metric m_s induces the SM_2 topology.

That will imply that the metric m_p induces the associated product topology. However, it turns out that the metric m_w does *not* induce the WM_2 topology. We will show that the WM_2 topology coincides with the product topology, so that the Hausdorff metric can be used to define the WM_2 topology via m_p in (10.5).

Closely paralleling the d or M_1 metrics, we have $m_p \leq m_s$ on $D([0, T], \mathbb{R}^k)$ and $m_p = m_w = m_s$ on $D([0, T], \mathbb{R}^1)$. Just as with d , we use m without subscript when the functions are real valued. Example ??, which showed that WM_1 is strictly weaker than SM_1 also shows that WM_2 is strictly weaker than SM_2 . Example ?? shows that the SM_2 topology is strictly weaker than the SM_1 topology.

Note that μ_s in (10.1) is *not* symmetric in its two arguments. Example 12.10.1 of the book shows that if $\mu_s(x, x_n) \rightarrow 0$ as $n \rightarrow \infty$, we need not have $\mu_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$.

6.10.1. The Hausdorff Metric Induces the SM_2 Topology

We now show that m_s induces the SM_2 topology.

Theorem 6.10.1. (the Hausdorff metric m_s induces the SM_2 topology)
If $\mu_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, then $\mu_s(x, x_n) \rightarrow 0$ as $n \rightarrow \infty$. Hence, $\mu_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ if and only if $m_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$.

Proof. Our proof will exploit lemmas below. Suppose that $\mu_s(x_n, x) \rightarrow 0$ but $\mu_s(x, x_n) \not\rightarrow 0$ as $n \rightarrow \infty$. Since $\mu_s(x, x_n) \not\rightarrow 0$, there exists $(z, t) \in \Gamma_x$ for which it is not possible to find $(z_n, t_n) \in \Gamma_{x_n}$ for $n \geq 1$ such that $(z_n, t_n) \rightarrow (z, t)$ as $n \rightarrow \infty$, but that contradicts Lemma 6.10.4 below. ■

In order to complete the proof of Theorem 6.10.1, we prove the following four lemmas.

Lemma 6.10.1. *Suppose that $\mu_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$. If $(z_n, t_n) \in \Gamma_{x_n}$ for $n \geq 1$, then there exists a subsequence $\{(z_{n_k}, t_{n_k})\}$ with $(z_{n_k}, t_{n_k}) \rightarrow (z, t)$ as $n_k \rightarrow \infty$ for some $(z, t) \in \Gamma_x$. Moreover, the limits of all convergent subsequences must be in Γ_x .*

Proof. Suppose that $\mu_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ and consider any sequence $\{(z_n, t_n)\}$ with $(z_n, t_n) \in \Gamma_{x_n}$ for $n \geq 1$. By the definition of μ_s , there must exist $(z'_n, t'_n) \in \Gamma_x$ such that $\|(z_n, t_n) - (z'_n, t'_n)\| \rightarrow 0$ as $n \rightarrow \infty$. Since Γ_x is compact, there exists a convergent subsequence of the sequence $\{(z'_n, t'_n)\}$; i.e., there exists $\{(z'_{n_k}, t'_{n_k})\}$ such that $(z'_{n_k}, t'_{n_k}) \rightarrow (z, t)$ for some $(z, t) \in \Gamma_x$. By the triangle inequality, we must also have $(z_{n_k}, t_{n_k}) \rightarrow (z, t)$ as $n_k \rightarrow \infty$. Finally, suppose (z_{n_k}, t_{n_k}) is an arbitrary convergent subsequence of $\{(z_n, t_n)\}$. By the argument above, there exists $(z, t) \in \Gamma_x$ such that a subsequence $(z_{n_{k_j}}, t_{n_{k_j}}) \rightarrow (z, t)$ as $n_{k_j} \rightarrow \infty$. This implies that (z, t) must be the limit of the convergent subsequence $\{(z_{n_k}, t_{n_k})\}$. ■

Lemma 6.10.2. *Suppose that $\mu_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, $t \notin \text{Disc}(x)$ and $(z_n, t) \in \Gamma_{x_n}$ for $n \geq 1$. Then $z_n \rightarrow x(t)$ as $n \rightarrow \infty$.*

Proof. By Lemma 6.10.1, there is a subsequence $(z_{n_k}, t) \rightarrow (z, t) \in \Gamma_x$, but $z = x(t)$ for $(z, t) \in \Gamma_x$ because $t \notin \text{Disc}(x)$. Since all convergent subsequences must have the same limit, $z_n \rightarrow z = x(t)$ as $n \rightarrow \infty$. ■

Corollary 6.10.1. *If $t \notin \text{Disc}(x)$ and $\mu_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, then $x_n(t) \rightarrow x(t)$ and $x_n(t-) \rightarrow x(t)$ in $\mathbb{R}^k$ as $n \rightarrow \infty$.*

Lemma 6.10.3. *If $\mu_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ and $(z, t) \in \Gamma_x$, then for any i , $1 \leq i \leq k$, there exist $(z_n, t_n) \in \Gamma_{x_n}$ for $n \geq 1$ such that $|z_n^i - z^i| \vee |t_n - t| \rightarrow 0$.*

Proof. The conclusion follows from Corollary 6.10.1 if $t \notin \text{Disc}(x)$, so suppose that $t \in \text{Disc}(x)$. Then z belongs to the segment $[x(t-), x(t)]$. First choose $t'_m > t$ with $t'_m \notin \text{Disc}(x)$ for all m and $t'_m \downarrow t$ as $m \rightarrow \infty$. By Lemma 6.10.2, there exist $(z'_{m,n}, t'_m) \in \Gamma_{x_n}$ such that $z'_{m,n} \rightarrow x(t'_m)$ as $n \rightarrow \infty$. Next choose $t''_m < t$ with $t''_m \notin \text{Disc}(x)$ for all m and $t''_m \uparrow t$ as $m \rightarrow \infty$. By Lemma 6.10.2 again, there exist $(z''_{m,n}, t''_m) \in \Gamma_{x_n}$ such that $z''_{m,n} \rightarrow x(t''_m)$ as $n \rightarrow \infty$. The diagonal sequences $(z'_{n,n}, t'_n)$ and $(z''_{n,n}, t''_n)$ thus belong to Γ_{x_n} and satisfy $t'_n \downarrow t$, $t''_n \uparrow t$, $z'_{n,n} \rightarrow x(t)$ and $z''_{n,n} \rightarrow x(t-)$ as $n \rightarrow \infty$. Since $\Gamma_{x_n}^i$ is a continuous real-valued curve, every value in the segment $[z_{n,n}^i, z_{n,n}^{i}]$ is realized for some $t_n^{i''}$ with $t_n^{i''} \leq t_n^{i''} \leq t'_n$. Hence, for any $(z, t) \in \Gamma_x$, there exists $(z_n^{i''}, t_n^{i''}) \in \Gamma_{x_n}$ such that $(z_n^{i''}, t_n^{i''}) \rightarrow (z^i, t)$ as $n \rightarrow \infty$. ■

Lemma 6.10.4. *If $\mu_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ and $(z, t) \in \Gamma_x$, then there exist $(z_n, t_n) \in \Gamma_{x_n}$ for $n \geq 1$ such that $\|(z_n, t_n) - (z, t)\| \rightarrow 0$ as $n \rightarrow \infty$.*

Proof. If $t \notin \text{Disc}(x)$, then we can take $(x_n(t), t) \in \Gamma_{x_n}$ or $(x_n(t-), t) \in \Gamma_{x_n}$ by Corollary 6.10.1. Hence it suffices to assume that $t \in \text{Disc}(x)$. Then, by the first part of the proof of Lemma 6.10.3, it suffices to consider (z, t) with $z \neq x(t)$ and $z \neq x(t-)$. For at least one coordinate i , either $x^i(t-) < z < x^i(t)$ or $x^i(t) > z > x^i(t)$. Consider one such coordinate. By Lemma 6.10.3, there is $(z_n, t_n) \in \Gamma_{x_n}$ such that $t_n \rightarrow t$ and $z_n^i \rightarrow z^i$ as $n \rightarrow \infty$. Moreover, since $\mu_s(x_n, x) \rightarrow 0$, given $(z_n, t_n) \in \Gamma_{x_n}$, we must have $(z'_n, t'_n) \in \Gamma_x$ such that $\|z_n - z'_n\| \vee |t_n - t'_n| \rightarrow 0$. Since $t_n \rightarrow t$, we must also have $t'_n \rightarrow t$. Since $z_n^i \rightarrow z^i$ and Γ_x contains the line joining $(x(t-), t)$ and $(x(t), t)$, we must have $z'_n \rightarrow z$ as well, which implies that $z_n \rightarrow z$, establishing the desired conclusion. ■

6.10.2. WM_2 is the Product Topology

We now observe that m_p induces the WM_2 topology.

Theorem 6.10.2. (WM_2 is the product topology) $\mu_w(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ for μ_w in (10.2) if and only if $m_p(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ for m_p in (10.5), so that the WM_2 topology on $D([0, T], \mathbb{R}^k)$ coincides with the product topology.

Proof. First, if $\mu_w(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, then $\mu_w(x_n^i, x^i) \rightarrow 0$ for each i , but $\mu_w(x_n^i, x^i) = \mu_s(x_n^i, x^i)$, so that $\mu_s(x_n^i, x^i) \rightarrow 0$ and $m_p(x_n, x) \rightarrow 0$ by Theorem 6.10.1. Conversely, suppose that $m_p(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$. Lemma 6.10.1 implies that $\cup_{n \geq 1} \Gamma_{x_n^i}$ is compact for each i , $1 \leq i \leq k$. That in turn implies that $\cup_{n \geq 1} G_{x_n}$ is compact. Hence, if $(z_n, t_n) \in G_{x_n}$ for $n \geq 1$, then every subsequence necessarily has a convergent subsubsequence. To have $\mu_w(x_n, x) \not\rightarrow 0$, we must have a subsequence of $\{(z_n, t_n)\}$ converge to a limit not in G_x . We will show that is not possible. Consider $(z_n, t_n) \in G_{x_n}$, $n \geq 1$. Since $t_n \in [0, T]$ for all n , there exists a subsequence (z_{n_k}, t_{n_k}) such that $t_{n_k} \rightarrow t$ for some t , $0 \leq t \leq T$. Since $m_p(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, there is a subsequence $\{(z_{n_{k_j}}, t_{n_{k_j}})\}$ such that $z_{n_{k_j}}^i \rightarrow z^i$ for some z^i where $(z^i, t) \in \Gamma_{x^i}$. Moreover, there are such subsequences for all i , $1 \leq i \leq k$, so that $z_n^i \rightarrow z^i$ for all i along the final subsequence. Moreover, $(z^i, t) \in \Gamma_{x^i}$ for all i , but this implies that $(z, t) \in G_x$. Hence every subsequence of (z_n, t_n) has a convergent subsubsequence and every convergent subsequence of $\{(z_n, t_n)\}$ has limit $(z, t) \in G_x$. That implies that $\mu_w(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$. ■

6.11. Alternative Characterizations of M_2 Convergence

We now give alternative characterizations of the SM_2 and WM_2 topologies.

6.11.1. M_2 Parametric Representations

We first observe that the SM_2 and WM_2 topologies can be defined just like the SM_1 and WM_1 topologies in Section 6.3. For this purpose, we say that a *strong M_2 (SM_2) parametric representation* of x is a continuous function (u, r) mapping $[0, 1]$ onto Γ_x such that r is nondecreasing. A *weak M_2 (WM_2) parametric representation* of x is a continuous function mapping $[0, 1]$ into G_x such that r is nondecreasing with $r(0) = 0$, $r(1) = T$ and $u(1) = x(T)$. The corresponding M_1 parametric representations are nondecreasing using the order defined on the graphs Γ_x and G_x in Section 2. In contrast, only the component function r is nondecreasing in the M_2 parametric representations. Let $\Pi_{s,2}(x)$ and $\Pi_{w,2}(x)$ be the sets of all SM_2 and WM_2 parametric representations of x .

Paralleling (3.7) and (3.8), define the distance functions

$$d_{s,2}(x_1, x_2) \equiv \inf_{\substack{(u_j, r_j) \in \Pi_{s,2}(x_j) \\ j=1,2}} \{ \|u_1 - u_2\| \vee \|r_1 - r_2\| \} \quad (11.1)$$

and

$$d_{w,2}(x_1, x_2) \equiv \inf_{\substack{(u_j, r_j) \in \Pi_{w,2}(x_j) \\ j=1,2}} \{ \|u_1 - u_2\| \vee \|r_1 - r_2\| \} . \quad (11.2)$$

We then can say that $x_n \rightarrow x$ as $n \rightarrow \infty$ for a sequence or net $\{x_n\}$ if $d_{s,2}(x_n, x) \rightarrow 0$ or $d_{w,2}(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$. A difficulty with this approach, just as for the WM_1 topology, is that neither $d_{s,2}$ nor $d_{w,2}$ is a metric.

6.11.2. SM_2 Convergence

We now establish the equivalence of several alternative characterizations of convergence in the SM_2 topology. To have a characterization involving the local behavior of the functions, we use the uniform-distance function $\bar{w}_s(x, x_2, t, \delta)$ in (4.6). We also use the related uniform-distance functions

$$\bar{w}_s(x_1, x_2, \delta) \equiv \sup_{0 \leq t \leq T} \bar{w}(x_1, x_2, t, \delta) . \quad (11.3)$$

$$\bar{w}_s^*(x_1, x_2, t, \delta) \equiv \|x_1(t) - [x_2((t - \delta) \vee 0), x_2((t + \delta) \wedge T)]\| \quad (11.4)$$

$$\bar{w}_s^*(x_1, x_2, \delta) \equiv \sup_{0 \leq t \leq T} \bar{w}_s^*(x_1, x_2, t, \delta) . \quad (11.5)$$

We now define new oscillation functions. The first is

$$\bar{w}_s^*(x, t, \delta) \equiv \sup\{\|x(t) - [x(t_1), x(t_2)]\|\} , \quad (11.6)$$

where the supremum is over

$$0 \vee (t - \delta) \leq t_1 \leq [0 \vee (t - \delta)] + \delta/2 \text{ and } [T \wedge (t + \delta)] - \delta/2 \leq t_2 \leq (t + \delta) \wedge T.$$

The second is

$$\bar{w}_s^*(x, \delta) \equiv \sup_{0 \leq t \leq T} \bar{w}_s^*(x, t, \delta) . \quad (11.7)$$

The uniform-distance function $\bar{w}_s^*(x_1, x_2, \delta)$ in (11.5) and the oscillation function $\bar{w}_s^*(x, \delta)$ in (11.7) were originally used by Skorohod (1956).

As before, T need not be a continuity point of x in $D([0, T], \mathbb{R}^k)$. Unlike for the M_1 topology, we can have $x_n \rightarrow x$ in (D, M_2) without having $x_n(T) \rightarrow x(T)$.

Let $v(x, A)$ represent the oscillation of x over the set A as in (2.5).

Theorem 6.11.1. (characterizations of SM_2 convergence) *The following are equivalent characterizations of $x_n \rightarrow x$ as $n \rightarrow \infty$ in (D, SM_2) :*

(i) $d_{s,2}(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ for $d_{s,2}$ in (11.1); i.e., for any $\epsilon > 0$ and n sufficiently large, there exist $(u, r) \in \Pi_{s,2}(x)$ and $(u_n, r_n) \in \Pi_{s,2}(x_n)$ such that $\|u_n - u\| \vee \|r_n - r\| < \epsilon$.

(ii) $m_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ for the metric m_s in (10.3).

(iii) $\mu_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ for μ_s in (10.1).

(iv) Given $\bar{w}_s(x_1, x_2, \delta)$ defined in (11.3),

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{w}_s(x_n, x, \delta) = 0 . \quad (11.8)$$

(v) For each t , $0 \leq t \leq T$,

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{w}_s(x_n, x, t, \delta) = 0 \quad (11.9)$$

for $\bar{w}_s(x_1, x_2, t, \delta)$ in (4.6).

(vi) For all $\epsilon > 0$ and all n sufficiently large, there exist finite ordered subsets A of Γ_x and A_n of Γ_{x_n} , as in (3.9) where $(z_1, t_1) \leq (z_2, t_2)$ if $t_1 \leq t_2$, of the same cardinality such that $\hat{d}(A, \Gamma_x) < \epsilon$, $\hat{d}(A_n, \Gamma_{x_n}) < \epsilon$ and $d^*(A, A_n) < \epsilon$ for $\hat{d}$ in (3.10) and d^* in (5.6).

(vii) Given $\bar{w}_s^*(x_1, x_2, \delta)$ defined in (11.5),

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{w}_s^*(x_n, x, \delta) = 0 .$$

(viii) $x_n(t) \rightarrow x(t)$ as $n \rightarrow \infty$ for each t in a dense subset of $[0, T]$ including 0 and

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{w}_s^*(x_n, \delta) = 0 \quad (11.10)$$

for $\bar{w}_s^*(x, \delta)$ in (11.7).

Proof. We already have shown the equivalence (ii) $\leftrightarrow$ (iii) in Theorem 11.10.1. (i) $\rightarrow$ (ii). Suppose that (i) holds with ϵ and n given. Since the parametric representations in $\Pi_{s,2}(x)$ map onto the graph Γ_x , for any $(z_n, t_n) \in \Gamma_{x_n}$, we can find $s \in [0, 1]$ such that $(u_n(s), r_n(s)) = (z_n, t_n)$. For that s , $(u(s), r(s)) = (z, t)$ for some $(z, t) \in \Gamma_x$ and

$$\|(z_n, t_n) - (z, t)\| \leq \|u_n - u\| \vee \|r_n - r\| < \epsilon . \quad (11.11)$$

By the same reasoning, for any $(z, t) \in \Gamma_x$, there exists $(z_n, t_n) \in \Gamma_{x_n}$ such that (11.11) holds.

(ii) $\rightarrow$ (v). For x , t and ϵ given, find δ such that $v(x, [t - \delta, t]) < \epsilon/2$ and $v(x, [t, t + \delta]) < \epsilon/2$ for v in (2.5). Then apply (ii) to find n_0 such that $m_s(x_n, x) < \eta \equiv (\epsilon \wedge \delta)/2$ for $n \geq n_0$. Then, for each t' with $0 \vee (t - \eta) \leq t' \leq (t + \eta) \wedge T$, there must exist $(\bar{z}, \bar{t}) \in \Gamma_x$ such that

$$\|(x_n(t'), t') - (\bar{z}, \bar{t})\| < \eta \quad \text{for } n \geq n_0 .$$

Since $|\bar{t} - t| \leq |\bar{t} - t'| + |t' - t| < 2\eta < \delta$,

$$\|(\bar{z}, \bar{t}) - [x(t-), x(t)]\| < \epsilon/2 .$$

Consequently, for $n \geq n_0$,

$$\|x_n(t') - [x(t-), x(t)]\| < \eta + \epsilon/2 < \epsilon .$$

Since t' was arbitrary,

$$w_s(x_n, x, t, \delta) < \epsilon .$$

(v) $\leftrightarrow$ (iv). Characterization (iv) clearly implies (v), so that it suffices to show that (v) implies (iv). We will show that if (iv) fails, then so does (v). Hence suppose that (iv) does not hold. Then there must exist $\epsilon > 0$, such that for any $\delta > 0$ there is a subsequence $\{n_k\}$ such that $n_k \rightarrow \infty$ and $\bar{w}_s(x_{n_k}, x, \delta) > \epsilon$ for all n_k . Hence, there is an associated sequence t_{n_k} such that

$$\bar{w}_s(x_{n_k}, x, t_{n_k}, \delta) > \epsilon/2$$

for all n_k . However, $\{t_{n_k}\}$ has a convergent subsequence $\{t_{n_{k_j}}\}$ with $t_{n_{k_j}} \rightarrow t$ as $n_{k_j} \rightarrow \infty$ for some t . Note that, if $z_n \in [x(t_n-), x(t_n)]$ for all n , where $t_n \rightarrow t$, and if $z_n \rightarrow z$, then necessarily $z \in [x(t-), x(t)]$. Hence,

$$\bar{w}_s(x_{n_{k_j}}, x, t, 2\delta) > \epsilon/2$$

for all sufficiently large n_{k_j} . That implies that (11.9) does not hold, so that (v) fails.

(iv) $\rightarrow$ (vi). We construct the desired finite subsets A of Γ_x and A_n of Γ_{x_n} by considering two kinds of points in Γ_x . For $\epsilon > 0$ given, we let A contain at least one point (z, t) for each $t \in \text{Disc}(x, \epsilon/2)$. The other points have $t \in \text{Disc}(x)^c$. We first construct A for t outside a finite union of neighborhoods of points in $\text{Disc}(x, \epsilon/2)$. We then construct A_n and finally we complete the definition of A by adding appropriate points (z, t) for $t \in \text{Disc}(x, \epsilon/2)$, which depend on A_n . Thus the set A ultimately depends upon A_n and thus upon x_n and n .

Let $t(A)$ denote the set of t for which there is at least one pair (z, t) from Γ_x in A . We first identify $t(A)$. We include $\text{Disc}(x, \epsilon/2)$ in $t(A)$. Use (11.8) to find an η and an n_0 such that $\bar{w}_s(x_n, x, \eta) < \epsilon/4$ for all $n \geq n_0$. Let $t_1 < \dots < t_m$ be the ordered set of points in $\text{Disc}(x, \epsilon/2) - \{T\}$; let $t_0 = 0$ and $t_{m+1} = T$. Use the existence of left and right limits for x to identify points, for $1 \leq i \leq m$, points t'_i and t''_i in $\text{Disc}(x)^c$ such that $t''_{i-1} < t'_i < t_i < t''_i < t'_{i+1}$, $|t_i - t'_i| < \eta$, $|t_i - t''_i| < \eta$, $v(x, [t'_i, t_i]) < \epsilon/4$ and $v(x, [t_i, t''_i]) < \epsilon/4$ for $v(x, B)$ in (2.5). We include these points t'_i and t''_i in $t(A)$. We also include in A points t''_0 and t'_{m+1} from $\text{Disc}(x)^c$ such that $t_0 = 0 < t''_0 < t'_1$, $t''_m < t'_{m+1} < t_{m+1} = T$, $v(x, [0, t''_0]) < \epsilon/4$ and $v(x, [t'_{m+1}, T]) < \epsilon/4$. We also include the points 0 and T in $t(A)$. Moreover, we include the points $(x(t'_i), t'_i)$, $(x(t''_i), t''_i)$, $(x(0), 0)$ and $(x(T), T)$ in A itself. (Except possibly for T , these are the only possibilities since $t'_i, t''_i, 0 \in \text{Disc}(x)^c$.) We next define A for t in the compact set

$$C \equiv [0, T] - \bigcup_{i=1}^m (t'_i, t''_i) - [0, t''_0] - (t'_{m+1}, T]. \quad (11.12)$$

The set C is a finite union of the closed intervals $[t''_i, t'_{i+1}]$, $0 \leq i \leq m-1$. For each t in C not a boundary point of one of these subintervals, it is possible to find t' and t'' in the same subinterval as t such that $t' < t < t''$, $|t - t'| < \eta/4$, $|t - t''| < \eta/4$ and $v(x, [t', t'']) < \epsilon/2$. (Recall that $C \subseteq \text{Disc}(x, \epsilon/2)^c$.) For the boundary points t'_i and t''_i , include intervals $(\bar{t}_i, t'_i]$ and $[t''_i, t^*_i)$ with the same properties; these intervals are open in the relative topology on C . Also include intervals $[0, t^*)$ and $(\bar{t}, T]$ with the same properties; these intervals again are open in the relative topology on C . These open intervals form an open cover of C . Since C is compact, there exists a finite subcover. We let $t(A)$ contain one point t in $\text{Disc}(x)^c$ from each subinterval in the finite subcover; we also put $(x(t), t)$ into A . Let the set A be ordered according to the time points; i.e., $(z_1, t_1) \leq (z_2, t_2)$ if $t_1 \leq t_2$. So far, A contains points $(x(t), t)$ for $t \in \text{Disc}(x)^c$, including the boundary points t'_i and t''_i of C . We have completed the definition of $t(A)$, which includes $\text{Disc}(x, \epsilon/2)$. If $\{t_i\}$ is the ordered set of points in $t(A)$, then the construction above implies that $|t_{i+1} - t_i| < \eta$ for all i (where η has been chosen so that $\bar{w}_s(x_n, x, \eta) < \epsilon/4$).

We now construct the set A_n . By Theorem 11.4.1, condition (11.8) implies that $x_n(t) \rightarrow x(t)$ for each $t \in \text{Disc}(x)^c$. For each $t \in t(A) - \text{Disc}(x, \epsilon/2)$, let $t \in t(A_n)$ and $(x_n(t), t) \in A_n$. Since each such t belongs to $\text{Disc}(x)^c$, there is $n_1 \geq n_0$ such that $\|x_n(t) - x(t)\| < \epsilon/4$ for all $t \in t(A) - \text{Disc}(x, \epsilon/2)$ and for all $n \geq n_1$. Hence we have established $d^*(A, A_n) < \epsilon/4$ for $n \geq n_1$ over C (outside the neighborhoods of $\text{Disc}(x, \epsilon/2)$). We complete the definition of A_n by adding finitely many points (z, t) for t in the open interval (t'_i, t''_i) where t'_i and t''_i are the adjacent points in $t(A)$ to $t_i \in \text{Disc}(x, \epsilon/2)$. We also do this for the interval $(t'_{m+1}, T]$ if $T \in \text{Disc}(x, \epsilon/2)$. We do this for all $t_i \in \text{Disc}(x, \epsilon/2)$ so that overall $\hat{d}(A_n, \Gamma_{x_n}) < \epsilon/2$. This is always possible by Lemma 6.3.1. We next complete the definition of A by including a point (z, t_i) for each point (z, t) in A_n with $t \in (t'_i, t''_i)$. This ensures that A_n and A have the same cardinality. Since $d(A_n, \Gamma_{x_n}) \leq \epsilon/2$, $\bar{w}_s(x_n, x, \eta) < \epsilon/4$,

$$\|x_n(t'_i) - x(t'_i)\| < \epsilon/4, \|x_n(t''_i) - x(t''_i)\| < \epsilon/4,$$

$$\|x(t'_i) - x(t_i-)\| < \epsilon/4 \quad \text{and} \quad \|x(t''_i) - x(t_i)\| < \epsilon/4$$

for $n \geq n_1$, we can choose points in A so that $d^*(A_n, A) \leq \epsilon/2$ for $n \geq n_1$ and $\hat{d}(A, \Gamma_x) \leq \epsilon$, which completes the proof.

(vi) $\rightarrow$ (i). Suppose that ϵ is given and the sets A and A_n in (vi) have points (z_i, t_i) and $(z_{n,i}, t_{n,i})$, $0 \leq i \leq m$, where $t_0 = 0$ and $t_m = T$. Construct arbitrary parametric representations of (u, r) of x and (u_n, r_n) of x_n such

that

$$r(i/m) = t_i, \quad u(i/m) = z_i$$

and

$$r_n(i/m) = t_{n,i}, \quad u_n(i/m) = z_{n,i} .$$

Since $d^*(A_n, A) \leq \epsilon$,

$$\max_{0 \leq i \leq m} \{ |r(i/m) - r_n(i/m)| \vee \|u(i/m) - u_n(i/m)\| \} < \epsilon .$$

Since $\hat{d}(A, \Gamma_x) < \epsilon$ and $\hat{d}(A_n, \Gamma_{x_n}) < \epsilon$ too, by the triangle inequality,

$$\|r - r_n\| \vee \|u_n - u\| < 3\epsilon .$$

(iv) $\leftrightarrow$ (vii). Suppose that $0 \leq t \leq T$. If x is constant in the intervals $(0 \vee (t - 2\delta), t)$ and $[t, (t + 2\delta) \wedge T)$, then

$$[x(0 \vee (t' - \delta)), x((t' + \delta) \wedge T)] = [x(t-), x(t)]$$

for all t' with $0 \vee (t - \delta) < t' < (t + \delta) \wedge T$. Consequently, in that situation

$$\begin{aligned} & \sup_{0 \vee (t - \delta) < t' < (t + \delta) \wedge T} \{ \|x_n(t') - [x(0 \vee (t' - \delta)), x((t' + \delta) \wedge T)] \| \} \\ &= \sup_{0 \vee (t - \delta) < t' < (t + \delta) \wedge T} \{ \|x_n(t') - [x(t-), x(t)] \| \} . \end{aligned} \quad (11.13)$$

Thus if x is piecewise constant with the distance between successive discontinuities at least δ , then $\bar{w}_s^*(x_n, x, \delta/2) = \bar{w}_s(x_n, x, \delta/2)$. Hence, for ϵ given suppose that we can choose η to make $\bar{w}_s(x_n, x, \eta) < \epsilon/3$. Then approximate x by $x_c \in D_c$ such that $\|x - x_c\| < \epsilon/3$. For that x_c , let α be the minimum distance between successive discontinuities. Then, for $\delta < \eta \wedge (\alpha/2)$,

$$\begin{aligned} \bar{w}_s^*(x_n, x, \delta) &\leq \bar{w}_s^*(x_n, x_c, \delta) + \epsilon/3 \\ &\leq \bar{w}_s(x_n, x_c, \delta) + \epsilon/3 \\ &\leq \bar{w}_s(x_n, x, \eta) + 2\epsilon/3 \leq \epsilon . \end{aligned} \quad (11.14)$$

Alternatively, for ϵ given, suppose that we can choose η to make $\bar{w}_s^*(x_n, x, \eta) < \epsilon/3$. Following the same reasoning,

$$\begin{aligned} \bar{w}_s(x_n, x, \delta) &\leq \bar{w}_s(x_n, x_c, \delta) + \epsilon/3 \\ &\leq \bar{w}_s^*(x_n, x_c, \delta) + \epsilon/3 \\ &\leq \bar{w}_s^*(x_n, x, \eta) + 2\epsilon/3 \leq \epsilon . \end{aligned} \quad (11.15)$$

Hence (iv) is equivalent to (vii).

(ii)→(viii). By Theorem 11.4.1 and (ii)↔(v), (ii) implies that $x_n(t) \rightarrow x(t)$ for each $t \in \text{Disc}(x)^c$. It remains to show that (ii)→(11.10). For $\epsilon > 0$ given, first pick a piecewise-constant x_c such that $\|x - x_c\| \leq \epsilon/4$, which is possible by Lemma 6.3.1. Let γ be the $\mathbb{R}^k$ -valued function with $\gamma^i(t) = 1$, $0 \leq t \leq T$, $1 \leq i \leq k$. Then $x_c - (\epsilon/4)\gamma \leq x \leq x_c + (\epsilon/4)\gamma$, i.e.,

$$x_c(t) - \epsilon/4 \leq x(t) \leq x_c(t) + \epsilon/4 \quad \text{for } 0 \leq t \leq T.$$

Let the (α, β) -neighborhood of $x \in D$ be

$$N_{\alpha, \beta}(x) \equiv \{[x(t) - \alpha\gamma, x(t) + \alpha\gamma] \times [0 \vee (t - \beta), (t + \beta) \wedge T] : 0 \leq t \leq T\}. \quad (11.16)$$

Thus, $x \in N_{\epsilon/4, 0}(x_c)$ and $x_c \in N_{\epsilon/4, 0}(x)$. Now let α be the minimum distance between successive discontinuities in x_c , or to 0 or T for the leftmost and rightmost discontinuity points. Given (ii), choose n_0 so that $m_s(x_n, x) = \eta_n < \eta < (\epsilon \wedge \alpha)/4$ for $n \geq n_0$. Then $x_n \in N_{\eta + \epsilon/4, \eta}(x_c)$. Suppose that $\{t_i : 1 \leq i \leq m - 1\}$ is the set of discontinuities of x_c , with $t_0 = 0$ and $t_m = T$. By the construction above, the open intervals $(t_i - \eta, t_i + \eta)$ are disjoint, $1 \leq i \leq m - 1$. Now let $\delta = 2\eta$. Hence, if $t' \in (t_i - \eta, t_i + \eta)$ for $t_i \in \text{Disc}(x_c)$, then

$$t_{i-1} + \eta < t' - \delta < t' - \delta/2 < t_i - \eta < t_i + \eta < t' + \delta/2 < t' + \delta < t_{i+1} - \eta \quad (11.17)$$

for all i , $1 \leq i \leq m - 1$. On the other hand, if $t' \in [t_{i-1} + \eta, t_i - \eta] = B_{i,n}$, then necessarily either $(t' + \delta/2, t' + \delta)$ intersects $B_{i,n}$ or $(t' - \delta, t' - \delta/2)$ intersects $B_{i,n}$. Thus, for $n \geq n_0$ and each $t' \in [0, T]$, there exists $t_1 \in [0 \vee (t' - \delta), 0 \vee (t' - \delta) + \delta/2)$ and $t_2 \in (T \wedge (t' + \delta) - \delta/2, T \wedge (t' + \delta)]$ such that

$$\|x_n(t') - [x_n(t_1), x_n(t_2)]\| \leq 2((\epsilon/4) + \eta) < \epsilon; \quad (11.18)$$

i.e., $\bar{w}_s^*(x_n, \delta) < \epsilon$.

(viii)→(v). For x, t and ϵ given, choose η so that $0 < t - \eta < t < t + \eta < T$, $v(x, [t - \eta, t]) < \epsilon/4$ and $v(x, [t, t + \eta]) < \epsilon/4$. Now choose $\delta < \eta$ and $t' \in (t - \delta/2, t + \delta/2)$. For δ and t' given, find t_1, t_2 in $\text{Disc}(x)^c$ such that $t_1 < t < t_2$, $t' - \delta < t_1 < t' - \delta/2$ and $t' + \delta/2 < t_2 < t' + \delta$. Then choose n_0 so that $\|x_n(t_i) - x(t_i)\| < \epsilon/4$ for $i = 1, 2$ and $n \geq n_0$. Apply (viii) to choose $n_1 \geq n_0$ so that $\bar{w}_s^*(x_n, \delta) \leq \epsilon/2$. Then

$$\begin{aligned} \|x_n(t') - [x(t-), x(t)]\| &\leq \|x_n(t') - [x(t_1), x(t_2)]\| + \epsilon/4 \\ &\leq \|x_n(t') - [x_n(t_1), x_n(t_2)]\| + \epsilon/2 \\ &\leq \bar{w}_s^*(x_n, \delta) + \epsilon/2 \leq \epsilon \quad \text{for } n \geq n_1 \end{aligned} \quad (11.19)$$

Since t' is arbitrary in $(t - \delta/2, t + \delta/2)$,

$$\bar{w}_s(x_n, x, t, \delta/2) \leq \epsilon \quad \text{for } n \geq n_1 ,$$

which implies (v). ■

Remark 6.11.1. The equivalence (iii) $\leftrightarrow$ (vii) $\leftrightarrow$ (viii) was established by Skorohod (1956). ■

Remark 6.11.2. There is no analog to characterization (v) involving $\bar{w}_s^*(x_n, x, t, \delta)$ in (11.4) instead of $\bar{w}_s(x_n, x, t, \delta)$. For $t \in \text{Disc}(x)^c$,

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{w}_s^*(x_n, x, t, \delta) = 0$$

implies pointwise convergence $x_n(t) \rightarrow x(t)$, but not the local uniform convergence in Theorem 6.4.1. ■

6.11.3. WM_2 Convergence

Corresponding characterizations of WM_2 convergence follow from Theorem 6.11.1 because the WM_2 topology is the same as the product topology, by Theorem 6.10.2. Let

$$\bar{w}_w(x_1, x_2, \delta) \equiv \sup_{0 \leq t \leq T} \bar{w}_w(x_1, x_2, t, \delta) \quad (11.20)$$

for $\bar{w}_w(x_1, x_2, t, \delta)$ in (4.7).

Theorem 6.11.2. (characterizations of WM_2 convergence) *The following are equivalent characterizations of $x_n \rightarrow x$ as $n \rightarrow \infty$ in (D, WM_2) :*

(i) $d_{w,2}(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ for $d_{w,2}$ in (11.2); i.e., for any $\epsilon > 0$ and all n sufficiently large, there exist $(u, r) \in \Pi_{w,2}(x)$ and $(u_n, r_n) \in \Pi_{w,2}(x_n)$ such that $\|u_n - u\| \vee \|r_n - r\| < \epsilon$.

(ii) $m_p(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ for the metric m_p in (10.5).

(iii) Given $\bar{w}_w(x_1, x_2, \delta)$ defined in (11.20),

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{w}_w(x_n, x, \delta) = 0 .$$

(iv) For each t , $0 \leq t \leq T$,

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{w}_w(x_n, x, t, \delta) = 0 .$$

(v) For all $\epsilon > 0$ and all sufficiently large n , there exist finite ordered subsets A of G_x and A_n of Γ_{x_n} , of common cardinality m as in (3.9) with $(z_1, t_1) \leq (z_2, t_2)$ if $t_1 \leq t_2$, such that $\hat{d}(A, G_x) < \epsilon$, $\hat{d}(A_n, \Gamma_{x_n}) < \epsilon$ and $d^*(A, A_n) < \epsilon$ for all $n \geq n_0$, for $\hat{d}$ in (5.8) and d^* in (5.6).

Proof. (i)→(ii). Clearly, $d_{w,2}(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ implies that $d_{s,2}(x_n^i, x^i) \rightarrow 0$ as $n \rightarrow \infty$ for each i . By Theorem 6.11.1, that implies $m_s(x_n^i, x^i) \rightarrow 0$ as $n \rightarrow \infty$ for each i , which implies (ii).

(ii)↔(iii). By Theorem 6.11.1, (ii) is equivalent to

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{w}_s(x_n^i, x^i, \delta) = 0 \quad (11.21)$$

for each i , but that is equivalent to (iii) because

$$\max_{1 \leq i \leq k} \bar{w}_s(x_n^i, x^i, \delta) = \bar{w}_w(x_n, x, \delta) . \quad (11.22)$$

(iii)↔(iv). By Theorem 6.11.1, (iii) is equivalent to

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{w}_s(x_n^i, x^i, t, \delta) = 0 \quad (11.23)$$

for each i , but that is equivalent to (iv) because

$$\max_{1 \leq i \leq k} \bar{w}_s(x_n^i, x^i, t, \delta) = \bar{w}_w(x_n, x, t, \delta) . \quad (11.24)$$

(iii)→(v). Follow the proof of (iv)→(vi) in Theorem 6.11.1. Use (??) to find an η and an n_0 such that $\bar{w}_w(x_n, x, \eta) < \epsilon/4$ for all $n \geq n_0$. Define $t(A)$ as before, first by including $Disc(x, \epsilon/2)$ and then by adding points from $Disc(x)^c$ in the complement of the union of the intervals about the points in $Disc(x, \epsilon/2)$. Let A be defined for $t \in t(A) - Disc(x, \epsilon/2)$ just as before. Let A_n be defined just as before. We complete the definition of A by including a point (z_i, t_i) for each point (z, t) in A_n with $t \in (t'_i, t''_i)$. This ensures that A and A_n have the same cardinality. Since $d(A_n, \Gamma_{x_n}) \leq \epsilon/2$, $\bar{w}_w(x_n, x, \eta) < \epsilon/4$, $\|x_n(t'_i) - x(t'_i)\| < \epsilon/4$, $\|x_n(t''_i) - x(t''_i)\| < \epsilon/4$, $\|x(t'_i) - x(t'_i -)\| < \epsilon/4$ and $\|x(t''_i) - x(t_i)\| < \epsilon/4$ for $n \geq n_1$, we can choose these points to add to A so that $d^*(A_n, A) \leq \epsilon/2$ for $n \geq n_1$ and $\hat{d}(A, G_x) \leq \epsilon$. (Unlike in the proof of Theorem 6.11.1, here we cannot conclude that $\hat{d}(A, \Gamma_x) \leq \epsilon$.)

(v)→(i). Paralleling the proof of (v)→(i) in Theorem 11.5.2, suppose that the conditions of (v) hold and A, A_n and ϵ are given. Let (u, r) and (u_n, r_n)

be parametric representations of x and x_n such that

$$\begin{aligned} u(i/m) &= z_i, \quad r(i/m) = t_i \quad \text{for } (z_i, t_i) \in A \\ u_n(i/m) &= z_{n,i}, \quad r_n(i/m) = t_{n,i} \quad \text{for } (z_{n,i}, t_{n,i}) \in A_n. \end{aligned}$$

For any $s \in [0, 1]$ there is i such that $s_i \leq s \leq s_{i+1}$ and

$$\begin{aligned} &\|u_n(s) - u(s)\| \vee \|r_n(s) - r(s)\| \leq \|(u_n(s), r_n(s)) - u_n(s_i), r_n(s_i)\| \\ &+ \|(u_n(s_i), r_n(s_i)) - (u(s_i), r(s_i))\| + \|(u(s_i), r(s_i)) - (u(s), r(s))\| \\ &\leq \hat{d}(A_n, G_{x_n}) + d^*(A_n, A) + \hat{d}(A, G_x) \leq 3\epsilon. \quad \blacksquare \end{aligned}$$

Theorem 6.11.2 and Section 6.4 show that all forms of M convergence imply uniform convergence to continuous limit functions.

Corollary 6.11.1. (from WM_2 convergence to uniform convergence) *Suppose that $m_p(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$.*

(i) *If $t \in \text{Disc}(x)^c$, then*

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} v(x_n, x, t, \delta) = 0.$$

(ii) *If $x \in C$, then $\lim_{n \rightarrow \infty} \|x_n - x\| = 0$.*

Proof. For (i) combine Theorems 6.4.1 and 6.11.2. For (ii) add Lemma 6.4.2. $\blacksquare$

Convergence in WM_2 has the advantage that jumps in the converging functions must be inherited by the limit function.

Corollary 6.11.2. (inheritance of jumps) *If $x_n \rightarrow x$ in (D, WM_2) , $t_n \rightarrow t$ in $[0, T]$ and $x_n^i(t_n) - x_n^i(t_n-) \geq c > 0$ for all n , then $x^i(t) - x^i(t-) \geq c$.*

Proof. Apply Theorem 6.11.2 (iv). $\blacksquare$

Let $J(x)$ be the maximum magnitude (absolute value) of the jumps of the function x in D . We apply Corollary 8.5.1 to show that J is upper semicontinuous.

Corollary 6.11.3. (upper semicontinuity of J) *If $x_n \rightarrow x$ in (D, M_2) , then*

$$\overline{\lim}_{n \rightarrow \infty} J(x_n) \leq J(x).$$

Proof. Suppose that $x_n \rightarrow x$ in (D, WM_2) and there exists a subsequence $\{x_{n_k}\}$ such that $J(x_{n_k}) \rightarrow c$. Then there exist further subsubsequences $\{x_{n_{k_j}}\}$ and $\{t_{n_{k_j}}\}$, and a coordinate i , such that $t_{n_{k_j}} \rightarrow t$ for some $t \in [0, T]$ and $|x_{n_{k_j}}^i(t_{n_{k_j}}) - x_{n_{k_j}}^i(t_{n_{k_j}} -)| \rightarrow c$. Then Corollary 8.5.1 implies that $|x^i(t) - x^i(t-)| \geq c$. ■

6.11.4. Additional Properties of M_2

We conclude this section by discussing additional properties of the M_2 topologies. First we note that there are direct M_2 analogs of the M_1 results in Theorems 6.6.1, 6.7.1, 6.7.2 and 6.7.3.

Theorem 6.11.3. (extending SM_2 convergence to product spaces) *Suppose that $m_s(x_n, x) \rightarrow 0$ in $D([0, T], \mathbb{R}^k)$ and $m_s(y_n, y) \rightarrow 0$ in $D([0, T], \mathbb{R}^l)$ as $n \rightarrow \infty$. If*

$$Disc(x) \cap Disc(y) = \phi,$$

then

$$m_s((x_n, y_n), (x, y)) \rightarrow 0 \text{ in } D([0, T], \mathbb{R}^{k+l}) \text{ as } n \rightarrow \infty.$$

Proof. We use characterization (v) in Theorem 6.11.1. Using the discontinuity condition, it is easy to show that (11.9) holds for $[(x_n, y_n), (x, y)]$ when it holds separately for $[x_n, x]$ and $[y_n, y]$, because i.e., at most one of the segments $[(x(t-), x(t))]$ and $[y(t-), y(t))]$ contains more than a single point. ■

Corollary 6.11.4. (from WM_2 convergence to SM_2 convergence when the limit is in D_1) *If $m_p(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ and $x \in D_1$, then $m_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$.*

Theorem 6.11.4. (Lipschitz property of linear functions of the coordinate functions) *For any $x_1, x_2 \in D([0, T], \mathbb{R}^k)$ and $\eta \in \mathbb{R}^k$,*

$$m(\eta x_1, \eta x_2) \leq (\|\eta\| \vee 1) m_s(x_1, x_2) .$$

Proof. For (??), the key property is that

$$\Gamma_{\eta x} = \{(\eta z, t) : (z, t) \in \Gamma_x\} .$$

It suffices to show that for all $\epsilon > 0$ and $(z'_1, t_1) \in \Gamma_{\eta x_1}$ there exists $(z'_2, t_2) \in \Gamma_{\eta x_2}$ such that

$$|z'_1 - z'_2| \vee |t_1 - t_2| \leq (\|\eta\| \vee 1)m_s(x_1, x_2) + \epsilon.$$

However, for $(z'_1, t_1) \in \Gamma_{\eta x_1}$, there exists $(z_1, t_1) \in \Gamma_{x_1}$ such that $\eta z_1 = z'_1$. Then choose $(z_2, t_2) \in \Gamma_{x_2}$ such that

$$\|z_1 - z_2\| \vee |t_1 - t_2| \leq m_s(x_1, x_2) + \epsilon$$

Let $(z'_2, t_2) = (\eta z_2, t_2)$. Then

$$|z'_1 - z'_2| = |\eta z_1 - \eta z_2| \leq \|\eta\| \|z_1 - z_2\|. \quad \blacksquare$$

We have an analog of Corollary 6.7.1 for the M_2 topology.

Corollary 6.11.5. (SM_2 -continuity of addition) *If $m_s(x_n, x) \rightarrow 0$ and $m_s(y_n, y) \rightarrow 0$ in $D([0, T], \mathbb{R}^k)$ and*

$$Disc(x) \cap Disc(y) = \phi,$$

then

$$m_s(x_n + y_n, x + y) \rightarrow 0 \text{ in } D([0, T], \mathbb{R}^k).$$

Proof. First apply Theorem 6.11.3 to get $m_s((x_n, y_n), (x, y)) \rightarrow 0$ in $D([0, T], \mathbb{R}^{k+l})$. Then apply Theorem 6.11.4. $\blacksquare$

Theorem 6.11.5. (characterization of SM_2 convergence by convergence of all linear functions of the coordinates) *There is convergence $x_n \rightarrow x$ in $D([0, T], \mathbb{R}^k)$ as $n \rightarrow \infty$ in the SM_2 topology if and only if $\eta x_n \rightarrow \eta x$ in $D([0, T], \mathbb{R}^1)$ as $n \rightarrow \infty$ in the M_2 topology for all $\eta \in \mathbb{R}^k$.*

Proof. One direction is covered by Theorem 6.11.4. Suppose that $x_n \not\rightarrow x$ as $n \rightarrow \infty$ in SM_2 . Then apply part (v) of Theorem 6.11.1 to deduce that $\eta x_n \not\rightarrow \eta x$ as $\eta \rightarrow \infty$ for some η . Note that $\|a\| > 0$ for $a \in \mathbb{R}^k$ if and only if $|\eta a| > 0$ in $\mathbb{R}$ for some $\eta \in \mathbb{R}^k$. Also, $\|a - A\| > 0$ for $A \subseteq \mathbb{R}^k$ if and only if $|\eta a - \eta A| > 0$ in $\mathbb{R}$ for some $\eta \in \mathbb{R}^k$, where $\eta A = \{\eta b : b \in A\}$. $\blacksquare$

Just as with the M_1 topology, we can get convergence of sums under more general conditions than in Corollary 6.11.5. It suffices to have the jumps of x^i and y^i have common sign for all i . We can express this property by the condition (7.2).

Theorem 6.11.6. (continuity of addition at limits with jumps of common sign) *If $x_n \rightarrow x$ and $y_n \rightarrow y$ in $D([0, T], \mathbb{R}^k, SM_2)$ and if condition (7.2) holds, then*

$$x_n + y_n \rightarrow x + y \quad \text{in} \quad D([0, T], \mathbb{R}^k, SM_2) .$$

Proof. Apply the characterization of SM_2 convergence in Theorem 6.11.1 (v). At points t in $Disc(x)^c \cap Disc(y)^c$, use the local uniform convergence in Lemma 12.5.1 of the book and Corollary 6.11.1 here. For other t not in $Disc(x) \cap Disc(y)$, use Theorem 6.11.3. For $t \in Disc(x) \cap Disc(y)$, exploit condition (7.2) to deduce that, for all $\epsilon > 0$, there exists δ and n_0 such that

$$\bar{w}_s(x_n + y_n, x + y, t, \delta) \leq w_s(x_n, x, t, \delta) + w_s(y_n, y, t, \delta) + \epsilon \quad (11.25)$$

for all $n \geq n_0$. ■

We now apply Theorem 6.11.5 to extend a characterization of convergence due to Skorohod (1956) to $\mathbb{R}^k$ -valued functions. For each $x \in D([0, T], \mathbb{R}^1)$ and $0 \leq t_1 < t_2 \leq T$, let

$$M_{t_1, t_2}(x) \equiv \sup_{t_1 \leq t \leq t_2} x(t) . \quad (11.26)$$

The proof exploits the SM_2 analog of Corollary 6.9.1.

In preparation for the next result, we state a basic lemma about preservation of convergence under restriction maps. For $x \in D([0, T], \mathbb{R}^k)$ and $0 \leq t_1 < t_2 \leq T^*$, let $r_{t_1, t_2} : D([0, T], \mathbb{R}^k) \rightarrow D([t_1, t_2], \mathbb{R}^k)$ be the restriction map, defined by $r_{t_1, t_2}(x)(s) = x(s)$, $t_1 \leq s \leq t_2$. We omit the proof.

Lemma 6.11.1. (continuity of restriction maps) *If $x_n \rightarrow x$ as $n \rightarrow \infty$ in $D([0, T], \mathbb{R}^k)$ with one of the SM_1 , WM_1 , SM_2 and WM_2 topologies and if $t_1, t_2 \in Disc(x)^c$, then*

$$r_{t_1, t_2}(x_n) \rightarrow r_{t_1, t_2}(x) \quad \text{as} \quad n \rightarrow \infty \quad \text{in} \quad D([t_1, t_2], \mathbb{R}^k)$$

with the same topology.

Theorem 6.11.7. (characterization of SM_2 convergence in terms of convergence of local extrema) *There is convergence $m_s(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ in $D([0, T], \mathbb{R}^k)$ if and only if*

$$M_{t_1, t_2}(\eta x_n) \rightarrow M_{t_1, t_2}(\eta x) \quad \text{as} \quad n \rightarrow \infty \quad (11.27)$$

for all $\eta \in \mathbb{R}^k$ and all points $t_1, t_2 \in \{T\} \cup Disc(x)^c$ with $t_1 < t_2$.

Proof. By Theorem 6.11.5, it suffices to consider the case of real-valued functions. By considering $\eta = \pm 1$ in (11.27), we get both the minimum and the maximum over $[t_1, t_2]$. It is easy to see that (11.27) for $\eta = \pm 1$ implies characterization (v) in Theorem 6.11.1: For x, t and ϵ given, choose γ so that $v(x, [t - \gamma, t]) < \epsilon/2$, $v(x, [t, t + \gamma]) < \epsilon/2$ and $0 < t - \gamma < t + \gamma < T$. Then finding n_0 such that $|M_{t_1, t_2}(\eta x_n) - M_{t_1, t_2}(\eta x)| < \epsilon/2$ for $n \geq n_0$, $\eta = \pm 1$ and

$$t - \gamma < t_1 < t - \delta < t < t + \delta < t_2 < t + \gamma$$

implies that $\bar{w}_s(x_n, x, t, \delta) < \epsilon$ for $n \geq n_0$. On the other hand, if $x_n \rightarrow x$ in $D([0, T], \mathbb{R}^1, M_2)$, then the restrictions converge in $D([t_1, t_2], \mathbb{R}^1, M_2)$ for all $t_1, t_2 \in \text{Disc}(x)^c$ by Lemma 6.11.1. If $m_s(x_n, x) < \epsilon$ in $D([t_1, t_2], \mathbb{R}^1, M_2)$, then clearly $|M_{t_1, t_2}(x_n) - M_{t_1, t_2}(x)| < \epsilon$ and $|M_{t_1, t_2}(-x_n) - M_{t_1, t_2}(-x)| < \epsilon$, so characterization (ii) of Theorem 6.11.1 implies (11.27). ■

We can apply the characterization of M_2 convergence in Theorem 6.11.7 to show the preservation of convergence under bounding functions in the M_2 topology. See Corollary 12.11.6 in the book.

6.12. Compactness

We have nothing to add in this final section.

Chapter 7

Useful Functions

7.1. Introduction

This chapter contains proofs omitted from Chapter 13 of the book, with the same title. As before, the theorems to be proved are restated here. The section and theorem numbers parallel Chapter 13 in the book, so that the proofs should be easy to find.

We consider four basic functions introduced in Section 3.5 of the book: composition, supremum, reflection and inverse. Another basic function is addition, but it has already been treated in Sections 12.6, 12.7 and 12.11 of the book. Our treatment of useful functions follows Whitt (1980), but the emphasis there was on the J_1 topology, even though the M_1 topology was used in places. In contrast, here the emphasis is on the M_1 and M_2 topologies.

Here is how this chapter is organized: We start in Section 7.2 by considering the composition map, which plays an important role in establishing FCLTs involving a random time change. We consider composition without centering in Section 7.2; then we consider composition with centering in Section 7.3.

In Section 7.4 we study the supremum function, both with and without centering. In Section 7.5 we apply the supremum results to treat the (one-sided one-dimensional) reflection map, which arises in queueing applications.

We start studying the inverse function in Section 7.6. We study the inverse map without centering in Section 7.6 and with centering in Section 7.7. In Section 7.8 we apply the results for inverse functions to obtain corresponding results for closely related counting functions.

In Section 7.9 we apply the previously established convergence-preservation results for the composition and inverse maps to establish stochastic-process

limits for renewal-reward stochastic processes. When the times between the renewals in the renewal counting process have a heavy-tailed distribution, we need the M_1 topology.

In Chapter 3 of the Internet Supplement we discuss pointwise convergence and its preservation under mappings. The preservation of pointwise convergence focuses on relations for individual sample paths, as in the queueing book by El-Taha and Stidham (1999). There we see that a function-space setting is not required for all convergence preservation.

7.2. Composition

This section is devoted to the composition function, mapping (x, y) into $x \circ y$, where

$$(x \circ y)(t) \equiv x(y(t)) \quad \text{for all } t.$$

The composition map is useful to treat random sums and, more generally, processes modified by a random time change; e.g., see Section 13.9 of the book on renewal-reward processes.

Henceforth in this chapter, unless stipulated otherwise, when $D \equiv D^k$, so that the range of functions is $\mathbb{R}^k$, we let D be endowed with the strong version of the J_1 , M_1 or M_2 topology, and simply write J_1 , M_1 or M_2 . It will be evident that most results also hold with the corresponding weaker product topology.

7.2.1. Preliminary Results

To ensure that $x \circ y \in D$, we will assume that y is also nondecreasing. We begin by defining subsets of $D \equiv D^k \equiv D([0, \infty), \mathbb{R}^k)$ that we will consider. Let D_0 be the subset of all $x \in D$ with $x^i(0) \geq 0$ for all i . Let $D_\uparrow$ and $D_{\uparrow\uparrow}$ be the subsets of functions in D_0 that are nondecreasing and strictly increasing in each coordinate. Let D_m be the subset of functions x in D_0 for which the coordinate functions x^i are monotone (either increasing or decreasing) for each i . Let C_0 , $C_\uparrow$, $C_{\uparrow\uparrow}$ and C_m be the corresponding subsets of C ; i.e., $C_0 \equiv C \cap D_0$, $C_\uparrow \equiv C \cap D_\uparrow$, $C_{\uparrow\uparrow} = C \cap D_{\uparrow\uparrow}$, and $C_m = C \cap D_m$.

It is important that all of these subsets are measurable subsets of D with the Borel σ -fields associated with the non-uniform Skorohod topologies, which all coincide with the Kolmogorov σ -field generated by the projection maps; see Theorems 11.5.2 and 11.5.3 in the book.

Returning to the composition map, we state the condition for $x \circ y \in D$ as a lemma.

Lemma 7.2.1. (criterion for $x \circ y$ to be in D) *For each $x \in D([0, \infty), \mathbb{R}^k)$ and $y \in D_{\uparrow}([0, \infty), \mathbb{R}_+)$, $x \circ y \in D([0, \infty), \mathbb{R}^k)$.*

A basic result, from pp. 145, 232 of Billingsley (1968), is the following. The continuity part involves the topology of uniform convergence on compact intervals.

Theorem 7.2.1. (continuity of composition at continuous limits) *The composition map from $D^k \times D_{\uparrow}^1$ to D^k is measurable and continuous at $(x, y) \in C^k \times C_{\uparrow}^1$.*

Our goal now is to obtain additional positive continuity results under extra conditions. We use the following elementary lemma.

Lemma 7.2.2. *If $y(t) \in \text{Disc}(x)$ and y is strictly increasing and continuous at t , then $t \in \text{Disc}(x \circ y)$.*

The following is the J_1 result.

Theorem 7.2.2. (J_1 -continuity of composition) *The composition map from $D^k \times D_{\uparrow}^1$ to D^k taking (x, y) into $(x \circ y)$ is continuous at $(x, y) \in (C^k \times D_{\uparrow}^1) \cup (D^k \times C_{\uparrow\uparrow}^1)$ using the J_1 topology throughout.*

Proof. First suppose that $(x_n, y_n) \rightarrow (x, y)$ in $D^k \times D_{\uparrow}^1$ with $(x, y) \in C^k \times D_{\uparrow}$. Choose $t_1 \in \text{Disc}(y)^c$. Then $y_n \rightarrow y$ for the restrictions to $[0, t_1]$; i.e., there exist $\lambda_n \in \Lambda([0, t_1])$ such that $\|y_n - y \circ \lambda_n\|_{t_1} \vee \|\lambda_n - e\|_{t_1} \rightarrow 0$. Choose t_2 such that $y(t_1) \leq t_2$ and $y_n(t_1) \leq t_2$ for all $n \geq 1$. Since $x \in C^k$, $\|x_n - x\|_{t_2} \rightarrow 0$. By the triangle inequality,

$$\|x_n \circ y_n - x \circ y \circ \lambda_n\|_{t_1} \leq \|x_n \circ y_n - x \circ y_n\|_{t_1} + \|x \circ y_n - x \circ y \circ \lambda_n\|_{t_1}. \quad (2.1)$$

The first term on the right in (2.1) converges to 0 because $\|x_n - x\|_{t_2} \rightarrow 0$ and the range of y_n is contained in $[0, t_2]$. The second term on the right in (2.1) converges to 0 because x is uniformly continuous over $[0, t_2]$ and $\|y_n - y \circ \lambda_n\|_{t_1} \rightarrow 0$.

Next suppose that $(x_n, y_n) \rightarrow (x, y)$ in $D^k \times D_{\uparrow}^1$ with $(x, y) \in D \times C_{\uparrow\uparrow}$. By Lemma 7.2.2 below, $y(t) \in \text{Disc}(x)^c$ for each $t \in \text{Disc}(x \circ y)^c$. However, for each $t' \in \text{Disc}(x)^c$, we have local uniform convergence of x_n to x , i.e.,

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} v(x_n, x, t', \delta) = 0; \quad (2.2)$$

see Section 12.4 in the book. Since $y_n(t) \rightarrow y(t)$ as $n \rightarrow \infty$, as a consequence of (2.2), we have $(x_n \circ y_n)(t) \rightarrow (x \circ y)(t)$ for each $t \in \text{Disc}(x \circ y)^c$. Now we show that the closure of the sequence $\{x_n \circ y_n : n \geq 1\}$ is compact in the J_1 topology. Since $(x_n \circ y_n)(t) \rightarrow (x \circ y)(t)$ for t in a countable dense subset, all limits of convergent subsequences must coincide with $x \circ y$. Since all convergent subsequences have the same limit, compactness implies that the sequence itself must converge; i.e., $x_n \circ y_n \rightarrow x \circ y$ (J_1). Hence it suffices to show that the closure of $\{x_n \circ y_n\}$ is compact, for which we apply Theorem 14.4 of Billingsley (1968). For an arbitrary t_1 , choose $t_2 > y(t_1)$ with $t_2 \in \text{Disc}(x)^c$. Then, for all sufficiently large n , $y_n(t_1) < t_2$ and $x_n \rightarrow x$ for the restrictions in $D([0, t_2], \mathbb{R}^k)$. It is easy to see that condition (14.49) and (14.50) in Billingsley (1968) hold. First, (14.49) holds because

$$\sup_{\substack{0 \leq s \leq t_1 \\ n \geq 1}} \|x_n \circ y_n(s)\| \leq \sup_{\substack{0 \leq s \leq t_2 \\ n \geq 1}} \|x_n\| < \infty, \quad (2.3)$$

since $x_n \rightarrow x$ in $D([0, t_2], \mathbb{R}^k, J_1)$. Next (14.50) holds because the oscillation functions for $x_n \circ y_n$ over $[0, t_1]$ be bounded above by the oscillation functions of x_n over $[0, t_2]$; e.g., since $y \in C_{\uparrow\uparrow}$ and $\|y_n - y\|_{t_1} \rightarrow 0$, for any δ_2 there exists n_0 and δ_1 such that $w''_{x_n \circ y_n}(\delta_1) \leq w''_{x_n}(\delta_2)$ for all $n \geq n_0$. ■

7.2.2. M -Topology Results

We have a different result for the M topologies.

Theorem 7.2.3. (*M -continuity of composition*) *If $(x_n, y_n) \rightarrow (x, y)$ in $D^k \times D_{\uparrow}^1$ and $(x, y) \in (D^k \times C_{\uparrow\uparrow}^1) \cup (C_m^k \times D_{\uparrow}^1)$, then $x_n \circ y_n \rightarrow x \circ y$ in D^k , where the topology throughout is M_1 or M_2 .*

In most applications we have $(x, y) \in D^k \times C_{\uparrow\uparrow}^1$, as is illustrated by the next section. That part of the M conditions is the same as for J_1 . The mode of convergence in Theorem 7.2.3 for $y_n \rightarrow y$ does not matter, because on $D_{\uparrow}^1$, convergence in the M_1 and M_2 topologies coincides with pointwise convergence on a dense subset of $[0, \infty)$, including 0; see Corollary ??.

It is easy to see that composition cannot in general yield convergence in a stronger topology, because $x \circ y = x$ and $x_n \circ y_n = x_n$, $n \geq 1$, when $y_n = y = e$, where $e(t) = t$, $t \geq 0$. Unlike for the J_1 topology, the composition map is in general *not* continuous at $(x, y) \in C \times D_{\uparrow}^1$ in the M topologies.

We actually prove a more general continuity result, which covers Theorem 7.2.3 as a special case.

Theorem 7.2.4. (more general M -continuity of composition) *Suppose that $(x_n, y_n) \rightarrow (x, y)$ in $D^k \times D_{\uparrow}^1$. If (i) y is continuous and strictly increasing at t whenever $y(t) \in \text{Disc}(x)$ and (ii) x is monotone on $[y(t-), y(t)]$ and $y(t-), y(t) \notin \text{Disc}(x)$ whenever $t \in \text{Disc}(y)$, then $x_n \circ y_n \rightarrow x \circ y$ in D^k , where the topology throughout is M_1 or M_2 .*

Theorem 7.2.3 follows easily from Theorem 7.2.4: First, on $D^k \times C_{\uparrow}^1$, y is continuous, so only condition (i) need be considered; it is satisfied because y is continuous and strictly increasing everywhere. Second on $C_m^k \times D_{\uparrow}^1$, x is continuous so only condition (ii) need be considered; it is satisfied because x is monotone everywhere. Hence it suffices to prove Theorem 7.2.4, which is done in Section 1.8 of the Internet Supplement. The general idea in our proof of Theorem 7.2.4 is to work with the characterization of convergence using oscillation functions evaluated at single arguments, exploiting Theorems 6.5.1 (v), 6.5.2 (iv), 6.11.1 (v) and 6.11.2 (iv).

We obtain a stronger result (M_1 convergence of $x_n \circ y_n$ given only M_2 convergence of x_n) if we do not need to invoke condition (i) in Theorem 7.2.4. A sufficient condition is for x to be continuous.

Theorem 7.2.5. (obtaining SM_1 convergence from WM_2 convergence) *If the conditions of Theorem 7.2.4 hold with $y(t) \notin \text{Disc}(x)$ for all t , then $x_n \circ y_n \rightarrow x \circ y$ in (D^k, SM_1) even if $x_n \rightarrow x$ only in (D^k, WM_2) .*

Proof. Apply Lemmas 7.2.4, 7.2.5 and 7.2.8 below. ■

We prove Theorem 7.2.4 by identifying four different cases, with each either having $t \in \text{Disc}(x \circ y)$ or not.

Proof of Theorem 7.2.4. We will establish the appropriate characterization of convergence $x_n \circ y_n \rightarrow x \circ y$ at each t separately, using Theorems 12.5.1 (v), 12.5.2 (iv), 12.11.1 (v) and 12.11.2 (iv) in the book.

There are four cases to consider:

- (i) $t \notin \text{Disc}(y)$ and $y(t) \notin \text{Disc}(x)$, so that $t \notin \text{Disc}(x \circ y)$;
- (ii) $t \in \text{Disc}(y)$, $x(u) = x(y(t-)) = x(y(t))$ for all $u \in [y(t-), y(t)]$ and $y(t-), y(t) \notin \text{Disc}(x)$, under which $t \notin \text{Disc}(x \circ y)$;
- (iii) $t \in \text{Disc}(y)$, $x(y(t-)) \neq x(y(t))$, x is monotone on $[y(t-), y(t)]$ and $y(t-), y(t) \notin \text{Disc}(x)$, under which $t \in \text{Disc}(x \circ y)$;
- (iv) $y(t) \in \text{Disc}(x)$ and y is continuous and strictly increasing at t so that $t \in \text{Disc}(x \circ y)$.

In case (ii) we have $t \notin \text{Disc}(x \circ y)$ even though $t \in \text{Disc}(y)$. The regularity conditions in case (ii) follow from condition (ii); since $x(y(t-)) = x(y(t))$, monotonicity reduces to a constant value over the subinterval. Case (iii) differs from case (ii) by having $x(y(t-)) \neq x(y(t))$, which makes $t \notin \text{Disc}(x \circ y)$. The regularity conditions in case (iii) again follow from condition (ii). The regularity conditions in case (iv) when $y(t) \in \text{Disc}(x)$ follow from condition (i). We use Lemma 7.2.2 in case (iv). In each case we know whether or not $t \in \text{Disc}(x \circ y)$. The four cases are covered by subsequent lemmas as follows: Case (i) by Lemmas 7.2.3–7.2.4; case (ii) by Lemma 7.2.5; case (iii) by Lemmas 7.2.6–7.2.8; and case (iv) by Lemma 7.2.10. ■

We now establish several lemmas in order to complete the proof of Theorem 7.2.4. Throughout, we assume that (x_n, y_n) , $n \geq 1$, and (x, y) are elements of $D^k \times D_{\uparrow}^1$. Refer to Section 12.4 of the book for the oscillation functions.

Lemma 7.2.3. *If $v(y_n, y, t, \delta_1) \leq \delta_2$ in $D_{\uparrow}^1$, then*

$$u(x_n \circ y_n, x \circ y, t, \delta_1) \leq v(x_n, x, y(t), \delta_2) + \bar{v}(x \circ y, t, \delta_1)$$

for v in (12.4.2), u in (12.4.1) and $\bar{v}$ in (12.4.3), all in Section 12.4 of the book.

Proof. By the condition, $|y_n(t_1) - y(t)| \leq \delta_2$ provided that $0 \vee (t - \delta_1) < t_1 < (t + \delta_1) \wedge T$. Hence, for t_1 in that range,

$$\begin{aligned} \|(x_n \circ y_n)(t_1) - (x \circ y)(t_1)\| &\leq \|x_n(y_n(t_1) - x(y(t)))\| \\ &\quad + \|x(y(t)) - x(y(t_1))\| \\ &\leq v(x_n, x, y(t), \delta_2) + \bar{v}(x \circ y, t, \delta_1) . \quad \blacksquare \end{aligned}$$

Lemma 7.2.4. *If $t \notin \text{Disc}(y)$, $y(t) \notin \text{Disc}(x)$,*

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} v(y_n, y, t, \delta) = 0 \tag{2.4}$$

and

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} v(x_n, x, y(t), \delta) = 0 , \tag{2.5}$$

then

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} v(x_n \circ y_n, x \circ y, t, \delta) = 0 . \tag{2.6}$$

Proof. Since $t \notin \text{Disc}(y)$ and $y(t) \notin \text{Disc}(x)$, $t \notin \text{Disc}(x \circ y)$ and $\bar{v}(x \circ y, t, \delta_1) \rightarrow 0$ as $\delta_1 \rightarrow 0$. We apply Lemma 7.2.3: For $\epsilon > 0$ given, choose δ_2 and n_1 so that

$$v(x_n, x, y(t), \delta_2) < \epsilon/2 \quad \text{for } n \geq n_1 .$$

Then choose δ_1 and $n_2 \geq n_1$ so that $\bar{v}(x \circ y, t, \delta_1) < \epsilon/2$ and

$$v(y_n, y, t, \delta_1) \leq \delta_2 \quad \text{for } n \geq n_2 .$$

By Lemma 7.2.3,

$$u(x_n \circ y_n, x \circ y, t, \delta_1) \leq \epsilon \quad \text{for } n \geq n_2 .$$

Since ϵ was arbitrary, we have shown that

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} u(x_n \circ y_n, x \circ y, t, \delta) = 0 ,$$

which is equivalent to (2.6) by Theorem 12.4.1 in the book. ■

Recall the m_p is the product metric inducing the WM_2 topology.

Lemma 7.2.5. *Suppose that $t \in \text{Disc}(y)$ but $y(t) \notin \text{Disc}(x)$, $y(t-) \notin \text{Disc}(x)$ and $x(y(t)) = x(y(t-))$ so that $t \notin \text{Disc}(x \circ y)$, i.e., case (ii) in Theorem 7.2.4. If $m_p(y_n, y) \rightarrow 0$ and $m_p(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, and $x(u) = x(y(t))$ for all $u \in [y(t-), y(t)]$, then (2.6) holds.*

Proof. Since $u \notin \text{Disc}(x)$ for all $u \in [y(t-), y(t)]$ and $m_p(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$, for $\epsilon > 0$ given, we can choose δ_1 and n_0 so that

$$\sup_{0 \vee (y(t-) - \delta_1) \leq u \leq (y(t) + \delta_1) \wedge T} \{ \|x_n(u) - x(u)\| \} \leq \epsilon \quad (2.7)$$

for all $n \geq n_0$ by Lemma 12.4.2 in the book. Since $x(u) = x(y(t))$ for $y(t-) \leq u \leq y(t)$ and x is continuous at $y(t-)$ and $y(t)$, from (2.7) we can obtain δ_2 such that

$$\sup_{0 \vee (y(t-) - \delta_2) \leq u \leq (y(t) + \delta_2) \wedge T} \{ \|x_n(u) - x(y(t))\| \} \leq 2\epsilon \quad (2.8)$$

for $n \geq n_0$. By right continuity and the existence of left limits, we can choose $t_1 < t < t_2$ such that

$$y(t_1) < y(t-) < y(t_1) + \delta_2/2 , \quad (2.9)$$

$$y(t) < y(t_2) < y(t) + \delta_2/2 , \quad (2.10)$$

$$\|(x \circ y)(t_j) - (x \circ y)(t)\| < \epsilon , \quad (2.11)$$

and $t_j \notin \text{Disc}(y)$ for $j = 1, 2$. Applying (2.4), we can choose $\delta_3 > 0$ and $n_1 \geq n_0$ so that

$$v(y_n, y, t_j, \delta_3) < \delta_2/2 \quad (2.12)$$

for all $n \geq n_1$ and $j = 1, 2$. Combining (2.9)–(2.12), and using the monotonicity of y_n and y , we have for $0 \vee (t - \delta_3) \leq t', t'' \leq (t + \delta_3) \wedge T$, $\|y_n(t') - \{y(t-), y(t)\}\| < \delta_2$. Thus, by (2.8),

$$\|x_n \circ y_n(t') - x \circ y(t'')\| \leq \|x_n \circ y_n(t') - x \circ y(t)\| + \|x \circ y(t) - x \circ y(t'')\| \leq 3\epsilon .$$

Since ϵ was arbitrary, we have established (2.6). ■

We now turn to case (iii). We first show how we can exploit the monotonicity condition.

Lemma 7.2.6. (characterization of M_2 convergence at a monotone limit)
Suppose that x is monotone on $[a, b]$. Then $x_n \rightarrow x$ in $D([a, b], \mathbb{R}^k, WM_2)$ if and only if $x_n \rightarrow x$ pointwise on a dense subset of $[a, b]$ and

$$\lim_{n \rightarrow \infty} w^*(x_n, [a, b]) = 0 , \quad (2.13)$$

where

$$w^*(x, [a, b]) \equiv \sup_{a \leq t_1 \leq t_2 \leq t_3 \leq b} \{ \|x(t_2) - [x(t_1), x(t_3)]\| \} . \quad (2.14)$$

These imply that $x_n \rightarrow x$ as $n \rightarrow \infty$ in SM_1 as well.

Proof. Clearly $w_s(x, \delta) \leq w^*(x, [a, b])$ on $D([a, b], \mathbb{R}^k)$ for all $\delta > 0$, where

$$w_s(x, \delta) \equiv \sup_{a \leq t \leq b} w_s(x, t, \delta)$$

for $w_s(x, t, \delta)$ in equation (12.4.4) of the book, so that (2.13) plus the pointwise convergence implies that $x_n \rightarrow x$ as $n \rightarrow \infty$ in SM_1 , by the basic characterization of SM_1 convergence, which in turn implies convergence in WM_2 . To go the other way, suppose that $w^*(x_n, [a, b]) \not\rightarrow 0$ as $n \rightarrow \infty$. Then there exist $\epsilon > 0$ and subsequences $\{n_k\}$, $\{t_{n_k, j}\}$ for $j = 1, 2$ and 3 such that $n_k \rightarrow \infty$ and

$$\|x_{n_k}(t_{n_k, 2}) - [x_{n_k}(t_{n_k, 1}), x_{n_k}(t_{n_k, 3})]\| > \epsilon \quad (2.15)$$

for all n_k . There are thus further subsequences $\{n'_k\}$, $\{t'_{n_k,2}\}$ for $j = 1, 2$, and 3 so that $t'_{n_k,j} \rightarrow t_j$ as $n'_k \rightarrow \infty$ for each j , where $t_1 \leq t_2 \leq t_3$. Assuming that $x_n \rightarrow x$ as $n \rightarrow \infty$ in WM_2 , we have $x_{n'_k}(t'_{n_k,j}) \rightarrow [[x(t_j-), x(t_j)]]$ as $n'_k \rightarrow \infty$, by the characterization of WM_2 convergence. This, with (2.15) and the monotonicity of x , implies that

$$\max_{1 \leq i \leq k} \{ \|x^i(t_2-) - [x^i(t_1-), x^i(t_3)]\|, \quad \|x^i(t_2) - [x^i(t_1-), x^i(t_3)]\| > 0, \}$$

which is impossible because x^i is monotone for each i . Hence, (2.13) must hold when $x_n \rightarrow x$ as $n \rightarrow \infty$ in WM_2 . ■

We will also apply the following elementary lemma, for which we omit the proof. We use the oscillation functions w_s in (12.4.4) and $\bar{v}$ in (12.4.3) of the book.

Lemma 7.2.7. *If*

$$y(t-) - \delta_2 \leq y_n(t_1) \leq y_n(t_2) \leq y(t) + \delta_2$$

whenever $0 < t - \delta_1 \leq t_1 \leq t_2 \leq t + \delta_1$, then

$$w_s(x_n \circ y_n, t, \delta_1) \leq \bar{v}(x_n, y(t), \delta_2) + \bar{v}(x_n, y(t-), \delta_2) + w^*(x_n, [y(t-), y(t)])$$

for w^ in (2.14).*

We apply Lemmas 7.2.6 and 7.2.7 to establish the following.

Lemma 7.2.8. *In case (iii), with $t \in \text{Disc}(y)$, $y(t-)$, $y(t) \notin \text{Disc}(x)$ and x monotone on $[y(t-), y(t)]$, if $(x_n, y_n) \rightarrow (x, y)$ in $D^k(WM_2) \times D^1_{\uparrow}(WM_2)$, then*

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w_s(x_n \circ y_n, t, \delta) = 0.$$

Proof. For any $\delta_2 > 0$ given, we can find δ_1 so that

$$y(t-) - \delta_2/2 \leq y(t_1) \leq y(t_2) \leq y(t) + \delta_2/2$$

for $0 \vee (t - \delta_1) \leq t_1 \leq t_2 \leq t + \delta_1$. By choosing continuity points of y , we can choose $n_2 \geq n_1$ so that

$$y(t-) - \delta_2 \leq y_n(t_1) \leq y_n(t_2) \leq y(t) + \delta_2$$

for all $n \geq n_2$. Hence we can apply Lemmas 7.2.6 and 7.2.7. By Lemma 7.2.6, $w^*(x_n, [y(t-), y(t)]) \rightarrow 0$ as $n \rightarrow \infty$. Since $x_n \rightarrow x$ and $y(t-), y(t) \notin \text{Disc}(x)$,

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{v}(x_n, y(t), \delta) = 0$$

and

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} \bar{v}(x_n, y(t-), \delta) = 0 .$$

An application of Lemma 7.2.7 completes the proof. ■

We now turn to case (iv). We first establish a preliminary result of independent interest, but which we do not directly need.

Lemma 7.2.9. *Suppose that $m_p(x_n, x) \rightarrow 0$ in D and $m(y_n, y) \rightarrow 0$ in $D_{\uparrow}^1$, but that $y(t) \in \text{Disc}(x)$. If y is strictly increasing and continuous in a neighborhood of t , then $(x_n \circ y_n)(t'_n) \rightarrow (x \circ y)(t')$ for all t' in a dense subset of neighborhood of t and all sequences $\{t'_n\}$ with $t'_n \rightarrow t'$.*

Proof. In the neighborhood of $y(t)$, there are at most countably many discontinuities of x . Since y is strictly increasing and continuous in a neighborhood of t , y is invertible there. Hence, for suitably small δ_2 and all but countably many t' in $(t - \delta_2, t + \delta_2)$, we simultaneously have y continuous at t' and x continuous at $y(t')$. At all such t' , we have $y_n(t'_n) \rightarrow y(t')$ and $x_n(y_n(t')) \rightarrow x(y(t'))$ whenever $t'_n \rightarrow t'$, because m_p -convergence implies local uniform convergence at continuity points, by virtue of Theorem 12.4.1 in the book.

Corollary 7.2.1. *If y is strictly increasing and continuous whenever $y(t) \in \text{Disc}(x)$ and $(x_n, y_n) \rightarrow (x, y)$ in $D_{\uparrow}^1(M_1) \times D_{\uparrow}^1(M_1)$, then $x_n \circ y_n \rightarrow x \circ y$ in $D_{\uparrow}^1(M_1)$.*

Proof. By Lemma 7.2.6, M_1 convergence on $D_{\uparrow}^1$ coincides with pointwise convergence on a dense subset. Apply Lemma 7.2.9. ■

Lemma 7.2.10. *If $m(y_n, y) \rightarrow 0$ in $D_{\uparrow}^1$, where y is continuous and strictly increasing at t , then for any $\delta > 0$, we can find $\delta_1 > 0$ such that, for all n sufficiently large,*

$$\begin{aligned} w_s(x_n \circ y_n, t, \delta_1) &\leq w_s(x_n, y(t), \delta) , \\ w_w(x_n \circ y_n, t, \delta_1) &\leq w_w(x_n, y(t), \delta) , \end{aligned}$$

$$\begin{aligned}\bar{w}_s(x_n \circ y_n, x \circ y, t, \delta_1) &\leq \bar{w}_s(x_n, x, y(t), \delta) , \\ \bar{w}_w(x_n \circ y_n, x \circ y, t, \delta_1) &\leq \bar{w}_w(x_n, x, y(t), \delta) .\end{aligned}$$

Proof. Since y is continuous at t , we can find $t_1 < t < t_2$ such that y is continuous at t_1 and t_2 and $|y(t) - y(t_j)| < \delta/2$ for $j = 1, 2$. Since $y_n \rightarrow y$ we can find n_0 such that $|y_n(t_j) - y(t_j)| < \delta/2$ for $n \geq n_0$ and $j = 1, 2$. By the triangle inequality, $|y_n(t_j) - y(t)| < \delta$ for $n \geq n_0$ and $j = 1, 2$. Let $\delta_1 = \min\{|t - t_1|, |t - t_2|\}$. Since y_n and y are nondecreasing, $|y_n(t') - y(t)| < \delta$ whenever $|t' - t| < \delta_1$. Hence

$$w_s(x_n \circ y_n, t, \delta_1) \leq w_s(x_n, y(t), \delta)$$

and

$$w_w(x_n \circ y_n, t, \delta_1) \leq w_w(x_n, y(t), \delta) .$$

Moreover, since y is continuous and strictly increasing, $x(y(t)-) = x(y(t))$. Hence

$$\bar{w}_s(x_n \circ y_n, x \circ y, t, \delta_1) \leq \bar{w}_s(x_n, x, y(t), \delta)$$

and

$$\bar{w}_w(x_n \circ y_n, x \circ y, t, \delta_1) \leq \bar{w}_w(x_n, x, y(t), \delta) . \quad \blacksquare$$

7.3. Composition with Centering

This section considers the composition map with centering. Nothing was omitted from the book here.

7.4. Supremum

In this section we consider the supremum function, mapping $D \equiv D([0, T], \mathbb{R})$ into itself according to

$$x^\uparrow(t) = \sup_{0 \leq s \leq t} x(s), \quad 0 \leq t \leq T. \quad (4.1)$$

7.4.1. The Supremum without Centering

The following elementary result is stated without proof.

Theorem 7.4.1. (Lipschitz property of the supremum function) *For any $x_1, x_2 \in D([0, T], \mathbb{R})$,*

$$\begin{aligned} d_{J_1}(x_1^\uparrow, x_2^\uparrow) &\leq d_{J_1}(x_1, x_2) , \\ d_{M_1}(x_1^\uparrow, x_2^\uparrow) &\leq d_{M_1}(x_1, x_2) , \\ d_{M_2}(x_1^\uparrow, x_2^\uparrow) &\leq d_{M_2}(x_1, x_2) . \end{aligned}$$

The conclusion in Theorem 7.4.1 can be recast in terms of pointwise convergence: Since $x^\uparrow$ is nondecreasing, convergence $x_n^\uparrow \rightarrow x^\uparrow$ in the M topologies is equivalent to pointwise convergence at continuity points of $x^\uparrow$, because on $D_\uparrow$ the M_1 and M_2 topologies coincide with pointwise convergence on a dense subset of $\mathbb{R}_+$ including 0; see Corollary 12.5.1 in the book. Thus the M topologies have not contributed much so far. We obtain more useful convergence-preservation results for the supremum map with the M topologies when we combine supremum with centering. As before, let e be the identity map, i.e., $e(t) = t$, $0 \leq t \leq T$.

7.4.2. The Supremum with Centering

The following is the main result stated as Theorem 13.4.2 in the book. Our object here is to prove it.

Theorem 7.4.2. (convergence preservation with the supremum function and centering) *Suppose that $c_n(x_n - e) \rightarrow y$ as $n \rightarrow \infty$ in $D([0, T], \mathbb{R})$ with one of the topologies J_1 , M_1 or M_2 , where $c_n \rightarrow \infty$.*

- (a) *If the topology is M_1 or M_2 , then $c_n(x_n^\uparrow - e) \rightarrow y$ in the same topology.*
- (b) *If the topology is J_1 , then $c_n(x_n^\uparrow - e) \rightarrow y$ if and only if y has no negative jumps.*

Before proving Theorem 7.4.2, we establish some preliminary lemmas. We first give an alternative expression for the result, in the form of a continuous mapping theorem. Let $y_n \equiv c_n(x_n - e)$. Then $s_n(y_n) = c_n(x_n^\uparrow - e)$, where

$$s_n(y) \equiv (y + c_n e)^\uparrow - c_n e \quad \text{for } y \in D . \quad (4.2)$$

Thus the conclusion of Theorem 7.4.2 can be expressed as $s_n(y_n) \rightarrow s(y) \equiv y$ when $y_n \rightarrow y$, with the appropriate topology.

Note that, for $x \in D$ and s_n in (4.2), $s_n(x)$ cannot have any negative jumps. For any $x \in D$, we can characterize $s_n(x)$ as the majorant which decreases by at most slope c_n at any time; i.e.,

$$s_n(x) = \inf\{y \in D : y \geq x, y(t_2) - y(t_1) \geq -c_n(t_2 - t_1)\}, \quad (4.3)$$

where we allow $0 \leq t_1 < t_2 \leq T$.

Lemma 7.4.1. *For any $x \in D$, $s_n(x)$ defined by (4.2) satisfies (4.3).*

Proof. First note that $s_n(x) \geq x$. Next note that

$$\begin{aligned} s_n(x)(t_2) - s_n(x)(t_1) &= (x + c_n e)^\uparrow(t_2) - (x + c_n e)^\uparrow(t_1) - c_n(t_2 - t_1) \\ &\geq -c_n(t_2 - t_1). \end{aligned}$$

Finally, suppose that $y \geq x$ and $y(t_2) - y(t_1) \geq -c_n(t_2 - t_1)$ for all $0 \leq t_1 < t_2 < T$. Then $s_n(y) = y$. Since $y \geq x$, $s_n(y) \geq s_n(x)$. Hence $y \geq s_n(x)$. ■

We can also bound $s_n(x)$ above for sufficiently large n by another majorant. Let the *left-local-majorant* of $x \in ([0, T], \mathbb{R})$ be

$$s_l^\epsilon(x)(t) = \sup_{0 \vee (t-\epsilon) \leq s \leq t} x(s), \quad 0 \leq t \leq T. \quad (4.4)$$

It is obvious that $x \leq s_l^\epsilon(x)$ for all x and $\epsilon > 0$. Moreover $s_l^\epsilon(x)(t)$ is nonincreasing as $\epsilon \downarrow 0$. We now show that $s_l^\epsilon(x) \rightarrow x$ in (D, M_2) as $\epsilon \downarrow 0$.

Lemma 7.4.2. *For any $x \in D$ and $\epsilon > 0$, there exists $\delta > 0$ such that*

$$d_{M_2}(x, s_l^\delta(x)) \leq \epsilon. \quad (4.5)$$

Proof. First, for x and ϵ given, apply Theorem 12.2.2 in the book to choose $x_c \in D_c$ such that $\|x - x_c\| < \epsilon/3$. For x_c , it is evident that there exists δ with $0 < \delta < \epsilon/3$ such that

$$d_{M_2}(s_l^\delta(x_c), x_c) < \delta < \epsilon/3 \quad \text{and} \quad \|s_l^\delta(x_c) - s_l^\delta(x)\| < \epsilon/3.$$

Hence,

$$d_{M_2}(x, s_l^\delta(x)) \leq \|x - x_c\| + d_{M_2}(x_c, s_l^\delta(x_c)) + \|s_l^\delta(x_c) - s_l^\delta(x)\| < \epsilon \quad \blacksquare \quad (4.6)$$

We now show that $s_n(x) \rightarrow x$ as $n \rightarrow \infty$ in the M_2 topology, uniformly over a large class of functions x .

Lemma 7.4.3. *Let s_n be as in (4.2), where $c_n \rightarrow \infty$. For any M and $\epsilon > 0$, there is an n_0 such that*

$$d_{M_2}(s_n(x), x) < \epsilon, \quad n \geq n_0, \quad (4.7)$$

for all x with $\|x\| \leq M$.

Proof. Let ϵ , M and x be given with $\|x\| \leq M$. Apply Lemma 7.4.2 to find δ such that $m(s_l^\delta(x), x) < \delta < \epsilon$. Choose n_0 so that $c_n \delta > 2M$ for $n \geq n_0$. Then, for $n \geq n_0$,

$$x(s) + c_n s - c_n t \leq x(t) \quad (4.8)$$

for all s , $0 \leq s \leq t - \delta$, $0 \leq t \leq T$, because under those conditions

$$x(s) + c_n s - c_n t \leq M - c_n \delta \leq -M \leq x(t). \quad (4.9)$$

Hence, for $n \geq n_0$,

$$x \leq s_n(x) \leq s_l^\delta(x), \quad (4.10)$$

so that, by Lemma 7.4.2, $s_n(x)$ is contained in an $M_2 \epsilon$ -neighborhood of x ; i.e., (4.7) holds. ■

Next, for the J_1 results we need the following.

Lemma 7.4.4. *If $x \in D([0, T], \mathbb{R})$ and x has no negative jumps, then for any $\epsilon > 0$ there is a $\delta > 0$ such that*

$$v^-(x, \delta) \equiv \sup_{\substack{0 \leq t \leq T \\ \text{ov}(t-\delta) \leq t' \leq t}} \{x(t') - x(t)\} < \epsilon. \quad (4.11)$$

Proof. Under the condition, for any $\epsilon > 0$ and all $t \in (0, T]$, there is a $\delta(t)$ such that $0 < t - \delta(t) < t$ and

$$x(t') \leq x(t) + \epsilon \quad \text{for all } t' \in (t - \delta(t), t). \quad (4.12)$$

By the right continuity of x at 0, there is a $\delta(0)$ such that $\|x(t') - x(0)\| < \epsilon$ for $0 \leq t' \leq \delta(0)$. The intervals $[0, \delta(0))$, $(t - \delta(t), t)$, $0 < t \leq T$, form an open cover of the compact set $[0, T]$. Hence there is a finite subcover. Let the subcover be chosen (modified) so that each t is in at most two subintervals. Let δ be the minimum length of the overlapping intervals, i.e.,

$$\delta = \min_i \{|t_i + \delta(t_{i+1}) - t_{i+1}|\} \wedge \delta(0). \quad (4.13)$$

Then, if t is any point in $[0, T]$, it either belongs to the subinterval $[0, \delta(0))$ or it is at least δ away from the left endpoint of one of its subintervals. Hence property (4.11) holds for δ in (4.13). ■

Proof of Theorem 7.4.2. (a) We will show that $s_n(x_n) \rightarrow x$ whenever $x_n \rightarrow x$, for s_n in (4.2). First consider the M_2 topology. Let M be a constant so that $\|x\| \leq M/2$. Since $d_{M_2}(x_n, x) \rightarrow 0$, there is an n_0 such that $\|x_n\| \leq M$ for all $n \geq n_0$. By the condition and Lemma 7.4.3, for any $\epsilon > 0$ there is an $n_1 \geq n_0$ such that $d_{M_2}(x_n, x) < \epsilon/2$ and $d_{M_2}(s_n(x_n), x_n) < \epsilon/2$ for $n \geq n_1$. Hence, by the triangle inequality, for $n \geq n_1$,

$$d_{M_2}(s_n(x_n), x) \leq d_{M_2}(s_n(x_n), x_n) + d_{M_2}(x_n, x) < \epsilon .$$

Next consider the M_1 topology. Since M_1 convergence implies M_2 convergence, we have $d_{M_2}(s_n(x_n), x) \rightarrow 0$ by the proof above. It thus suffices to strengthen convergence from M_2 to M_1 . In particular, we can apply part (v) of Theorem 12.5.1 in the book. By Theorem 12.4.1 in the book, the M_2 convergence implies the local uniform convergence at continuity points in condition (12.5.4) in the book, so it only remains to establish the oscillation function limit at discontinuity points in condition (12.5.5) in the book; i.e.,

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w_s(s_n(x_n), t, \delta) = 0 . \quad (4.14)$$

We show that if (4.14) fails, then necessarily we cannot have

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w_s(x_n, t, \delta) = 0 , \quad (4.15)$$

so that $x_n \not\rightarrow x$ (M_1), which is a contradiction. If (4.14) fails, then there must exist $\epsilon > 0$, $\delta_k \downarrow 0$ and $n_k \uparrow \infty$ such that

$$w_s(s_{n_k}(x_{n_k}), t, \delta_k) > \epsilon \quad \text{for all } k . \quad (4.16)$$

Let $y_{n_k} = s_{n_k}(x_{n_k})$. Given (4.16), there are two cases: In the first case, there exist $t_{n_k,1}$, $t_{n_k,2}$ and $t_{n_k,3}$ such that

$$0 \vee (t - \delta_k) \leq t_{n_k,1} < t_{n_k,2} < t_{n_k,3} \leq (t + \delta_k) \wedge T , \quad (4.17)$$

$y_{n_k}(t_{n_k,2}) > y_{n_k}(t_{n_k,1}) + \epsilon$ and $y_{n_k}(t_{n_k,2}) > y_{n_k}(t_{n_k,3}) + \epsilon$. However, $y_{n_k}(t_{n_k,2}) > y_{n_k}(t_{n_k,1}) + \epsilon$ implies that there must exist $t'_{n_k,2}$ with $t_{n_k,1} < t'_{n_k,2} \leq t_{n_k,2}$ and $x_{n_k}(t'_{n_k,2}) \geq y_{n_k}(t_{n_k,2})$. Since $y_{n_k}(t_{n_k,1}) \geq x_{n_k}(t_{n_k,1})$ and $y_{n_k}(t_{n_k,3}) \geq x_{n_k}(t_{n_k,3})$, we then must have $w_s(x_{n_k}, t, \delta_k) > \epsilon$, which contradicts (4.15).

In the second case, there exist $t_{n_k,1}$, $t_{n_k,2}$ and $t_{n_k,3}$ such that (4.17) holds, $y_{n_k}(t_{n_k,2}) < y_{n_k}(t_{n_k,1}) - \epsilon$ and $y_{n_k}(t_{n_k,2}) < y_{n_k}(t_{n_k,3}) - \epsilon$. By the last inequality, there must exist $t'_{n_k,3}$ with $t_{n_k,2} < t'_{n_k,3} \leq t_{n_k,3}$ such that

$x_{n_k}(t'_{n_k,3}) \geq y_{n_k}(t_{n_k,3}) - \epsilon$. Since $x_n \leq y_n$, $x_{n_k}(t_{n_k,2}) \leq y_{n_k}(t_{n_k,2})$. Finally, since $\{x_{n_k}\}$ is uniformly bounded, there is δ'_k where $\delta'_k \downarrow 0$ as $k \rightarrow \infty$, and $t'_{n_k,1}$ with $0 \vee (t - (\delta_k + \delta'_k)) \leq t'_{n_k,1} \leq t_{n_k,1}$ with $x_{n_k}(t'_{n_k,1}) \geq y_{n_k}(t_{n_k,1})$. Hence, we must have

$$w_s(x_{n_k}, t, \delta_k + \delta'_k) > \epsilon \quad \text{for all } k. \quad (4.18)$$

Since $\delta_k + \delta'_k \downarrow 0$ as $k \rightarrow \infty$, (4.18) again contradicts (4.15) and thus $x_n \rightarrow x(M_1)$. Thus, $d_{M_1}(s_n(x_n), x) \rightarrow 0$ as claimed.

(b) We now turn to the J_1 result. Given $c_n(x_n - e) \rightarrow y(J_1)$, there exists $\lambda_n \in \Lambda$ such that $\|c_n(x_n - e) - y \circ \lambda_n\| \rightarrow 0$ as $n \rightarrow \infty$. We want to show that $\|c_n(x_n^\uparrow - e) - y \circ \lambda_n\| \rightarrow 0$. Since $x_n^\uparrow \geq x_n$, it suffices to show, for any $\epsilon > 0$, that there is n_1 such that

$$c_n x_n(s') - c_n s \leq y(\lambda_n(s)) + \epsilon \quad \text{for } 0 \leq s' \leq s \leq T \quad (4.19)$$

for $n \geq n_1$. Choose n_0 such that $\|c_n(x_n - e) - y \circ \lambda_n\| < \epsilon/2$ for $n \geq n_0$. From (4.19), we see that it suffices to show that there is $n_1 \geq n_0$ such that

$$y(\lambda_n(s')) \leq y(\lambda_n(s)) + c_n(s - s') + \epsilon/2 \quad \text{for } 0 \leq s' \leq s \leq T. \quad (4.20)$$

Since y has no negative jumps, we can apply Lemma 7.4.4 to conclude that there is a δ such that $v^-(y, \delta) < \epsilon/2$ for $v^-(y, \delta)$ in (4.11). Then choose $n_1 \geq n_0$ such that $\|\lambda_n - e\| < \delta$ and $c_n \delta \geq \|y\|$ for $n \geq n_1$, and we obtain (4.20). Finally, recall that the maximum negative jump function is continuous, e.g., see p. 301 of Jacod and Shiryaev (1987); i.e.,

$$J_-(x) \equiv \sup_{0 \leq t \leq 1} \{x(t-) - x(t)\}. \quad (4.21)$$

Clearly, $J_-(c_n(x_n^\uparrow - e)) = 0$, so that if $c_n(x_n^\uparrow - e) \rightarrow y(J_1)$, then y must have no negative jumps. ■

We now obtain joint convergence in the stronger topologies on $D([0, T], \mathbb{R}^2)$ under the condition that the limit function have no negative jumps.

Theorem 7.4.3. (criterion for joint convergence) *Suppose that $c_n(x_n - e) \rightarrow y$ as $n \rightarrow \infty$ in $D([0, T], \mathbb{R})$ with one of the J_1 , M_1 or M_2 topologies, where $c_n \rightarrow \infty$. If, in addition, y has no negative jumps, then*

$$c_n(x_n - e, x_n^\uparrow - e) \rightarrow (y, y) \quad \text{as } n \rightarrow \infty \quad (4.22)$$

in $D([0, T], \mathbb{R}^2)$ with the strong version of the same topology, i.e., with SJ_1 , SM_1 or SM_2 .

Proof. For the SM_1 and SM_2 topologies, we will work with parametric representations, using the parametric representation $((u, u), r)$ for (y, y) . Given that $(c_n(x_n - e) \rightarrow y$, there exist parametric representations $(u_n, r_n) \in \Pi_s(c_n(x_n - e))$ and $(u, r) \in \Pi(y)$ such that $\|u_n - u\| \vee \|r_n - r\| \rightarrow 0$ as $n \rightarrow \infty$. We construct the desired parametric representations from these. Note that $(c_n^{-1}u_n + r_n, r_n) \in \Pi(x_n)$ and $(u'_n, r_n) \in \Pi(c_n(x_n^\uparrow - e))$ for

$$u'_n = c_n((c_n^{-1}u_n + r_n)^\uparrow - r_n) = (u_n + c_nr_n)^\uparrow - c_nr_n. \quad (4.23)$$

Note that $x_n^\uparrow$ has the jumps up of x_n , while $x_n^\uparrow$ is continuous when x_n has a jump down. Thus $((u_n, u'_n), r_n) \in \Pi_s(y_n, y'_n)$ for $y_n \equiv c_n(x_n - e)$ and $y'_n \equiv c_n(x_n^\uparrow - e)$. Of course $((u, u), r) \in \Pi_s((y, y))$. Thus it remains to show that

$$\|(u_n, u'_n) - (u, u)\| \vee \|r_n - r\| \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (4.24)$$

Given that $\|u_n - u\| \vee \|r_n - r\| \rightarrow 0$, it suffices to show that $\|u'_n - u\| \rightarrow 0$. Clearly, $u'_n \geq u_n$ for all n , so that it suffices to show that, for all $\epsilon > 0$, there exist n_1 such that $u'_n(s) < u(s) + \epsilon$ for all $n \geq n_1$ and $s \in [0, 1]$. Equivalently, by (4.23), it suffices to show that

$$u_n(s') + c_n(r_n(s') - r_n(s)) < u(s) + \epsilon, \quad 0 \leq s' \leq s \leq 1, \quad (4.25)$$

for all $n \geq n_1$. However, if we assume that the limit y has no negative jumps, then Lemma 7.4.4 implies that there is a $\delta > 0$ such that

$$u(s') \leq u(s) + \epsilon/2 \quad (4.26)$$

for all s, s' with $0 \leq s' \leq s \leq 1$ and $r(s) - r(s') < \delta$. Choose n_0 so that

$$\|u_n - u\| \vee \|r_n - r\| \leq (\delta \wedge \epsilon)/4 \quad \text{for } n \geq n_0.$$

Choose $n_1 \geq n_0$ so that

$$c_n\delta/4 \geq 2\|x\| \quad \text{for } n \geq n_1. \quad (4.27)$$

There are two cases: (i) $r_n(s) - r_n(s') \leq \delta/4$ and (ii) $r_n(s) - r_n(s') > \delta/4$. In case (i), $r(s) - r(s') < \delta$, so that by (4.26)

$$u_n(s') + c_n(r_n(s') - r_n(s)) \leq u_n(s') \leq u(s') + \epsilon/4 \leq u(s) + \epsilon, \quad (4.28)$$

so that (4.25) holds. In case (ii), by (4.27),

$$\begin{aligned} u_n(s') + c_n(r_n(s') - r_n(s)) &\leq u(s') + \epsilon/2 - c_n\delta/4 \\ &\leq u(s) + 2\|u\| - c_n\delta/4 + \epsilon/2 \\ &\leq u(s) + \epsilon, \end{aligned} \quad (4.29)$$

so that again (4.25) holds. Turning to J_1 , we note that the result already follows from the proof of Theorem 7.4.2 (ii) because the same homeomorphisms $\lambda_n \in \Lambda$ were used for both $c_n(x_n - e) \rightarrow y$ and $c_n(x_n^\uparrow - e) \rightarrow y$. ■

Corollary 7.4.1. *Under the conditions of Theorem 7.4.3,*

$$\|c_n(x_n^\uparrow - x_n)\| \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Proof. Apply subtraction to get

$$c_n(x_n - x_n^\uparrow) = c_n(x_n - e) - c_n(x_n^\uparrow - e) \rightarrow x - x(M_2).$$

Since the limit is continuous, the convergence holds in the uniform topology. ■

We next give an elementary result about the supremum function when the centering is in the other direction, so that x_n must be rapidly decreasing. Convergence $x_n^\uparrow(t) \rightarrow x(0)$ as $n \rightarrow \infty$ is to be expected, but that conclusion can not be drawn if the M_2 convergence in the condition is replaced by pointwise convergence.

Theorem 7.4.4. (convergence preservation with the supremum function when the centering is in the other direction) *Suppose that $c_n \rightarrow \infty$ and $x_n + c_n e \rightarrow y$ in $D([0, T], \mathbb{R}, M_2)$. Then*

$$\|x_n^\uparrow - z(y)\| \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

where $z(y)(t) \equiv y(0)$, $0 \leq t \leq T$.

Proof. The assumed M_2 convergence implies local uniform convergence at the origin: For any $\epsilon > 0$, there is a δ and an n_0 such that

$$\sup_{0 \leq t \leq \delta} |x_n(t) + c_n t - y(0)| \leq v(x_n, y, 0, \delta) < \epsilon$$

for $n \geq n_0$, where $v(x_1, x_2, t, \delta)$ is the modulus of continuity in (4.2) in Section 6.4. Hence, $x_n(t) \leq y(0) + \epsilon$ for all t , $0 \leq t \leq \delta$, and $n \geq n_0$. Use the conditions to find $n_1 \geq n_0$ such that $\|x_n + c_n e\| \leq \|y\| + \epsilon$ and $c_n \delta > 2\|y\|$ for $n \geq n_1$. Then, for $t > \delta$ and $n \geq n_1$,

$$x_n(t) = -c_n \delta + x_n(t) + c_n \delta \leq -c_n \delta + \|x_n + c_n e\| \leq -c_n \delta + \|y\| + \epsilon \leq y(0) + \epsilon.$$

Hence, $x_n^\uparrow(t) \leq y(0) + \epsilon$ for all t , $0 \leq t \leq T$, and $n \geq n_1$. On the other hand, for all t , $x_n^\uparrow(t) \geq x_n(0) \rightarrow y(0)$ as $n \rightarrow \infty$. ■

7.5. One-Dimensional Reflection

Closely related to the supremum function is the one-dimensional (one-sided) reflection mapping, which we have used to construct queueing processes. Indeed, the reflection mapping can be defined in terms of the supremum mapping as

$$\phi(x) \equiv x + (-x \vee 0)^\uparrow ;$$

i.e.,

$$\phi(x)(t) = x(t) - (\inf\{x(s) : 0 \leq s \leq t\} \wedge 0) , \quad 0 \leq t \leq T , \quad (5.1)$$

as in equation (2.5) in Section 5.2 of the book.

The Lipschitz property for the supremum function with the uniform topology in Lemma ?? immediately implies a corresponding result for the reflection map ϕ in (5.1).

Unfortunately, however, the Lipschitz property for the reflection map ϕ with the uniform topology does not even imply continuity in all the Skorohod topologies. In particular, ϕ is not continuous in the M_2 topology.

We do obtain positive results with the J_1 and M_1 topologies. As before, let d_{J_1} and d_{M_1} be the metrics in equations 3.2 and 3.4 in Section 3.3 of the book. For the J_1 result, we use the following elementary lemma.

Lemma 7.5.1. *For any $x \in D$ and $\lambda \in \Lambda$,*

$$\phi(x) \circ \lambda = \phi(x \circ \lambda) .$$

For the M_1 result, we use the following lemma. A fundamental difficulty for treating the more general multidimensional reflection map is that Lemma 7.5.2 below does not extend to the multidimensional reflection map; see Chapter 8.

Lemma 7.5.2. (preservation of parametric representations under reflections) *For any $x \in D$, if $(u, r) \in \Pi(x)$, then $(\phi(u), r) \in \Pi(\phi(x))$.*

Proof. In book. ■

Theorem 7.5.1. (Lipschitz property with the J_1 and M_1 metrics) *For any $x_1, x_2 \in D([0, T], \mathbb{R})$,*

$$d_{J_1}(\phi(x_1), \phi(x_2)) \leq 2d_{J_1}(x_1, x_2)$$

and

$$d_{M_1}(\phi(x_1), \phi(x_2)) \leq 2d_{M_1}(x_1, x_2) ,$$

where ϕ is the reflection map in (5.1).

Proof. In book. ■

Theorem 7.5.1 covers the standard heavy-traffic regime for one single-server queue when $\rho = 1$, where ρ is the traffic intensity. The next result covers the other cases: $\rho < 1$ and $\rho > 1$. We use the following elementary lemma in the easy case of the uniform metric.

Lemma 7.5.3. *Let d be the metric for the U , J_1 , M_1 or M_2 topology. Let $x \vee a : D \rightarrow D$ be defined by*

$$(x \vee a)(t) \equiv x(t) \vee a, \quad 0 \leq t \leq T. \quad (5.2)$$

Then, for any $x_1, x_2 \in D$,

$$d(x \vee a(x_1), x \vee a(x_2)) \leq d(x_1, x_2) .$$

Theorem 7.5.2. (convergence preservation with centering) *Suppose that $x_n - c_n e \rightarrow y$ in $D([0, T], \mathbb{R})$ with the U , J_1 , M_1 or M_2 topology.*

(a) If $c_n \rightarrow +\infty$, then

$$\phi(x_n) - c_n e \rightarrow y + \gamma(y) \quad \text{as } n \rightarrow \infty \quad \text{in } D$$

with the same topology, where

$$\gamma(y)(t) \equiv (-y(0)) \vee 0 = -(y(0) \wedge 0), \quad 0 \leq t \leq T.$$

(b) If $c_n \rightarrow -\infty$, $y(0) \leq 0$ and y has no positive jumps, then

$$\|\phi(x_n) - 0e\| \rightarrow 0 \quad \text{as } n \rightarrow \infty \quad \text{in } D ,$$

where $e(t) = t$, $0 \leq t \leq T$.

Proof. (a) Note that

$$\phi(x_n) - c_n e = x_n - c_n e + (-x_n \vee 0)^\uparrow ,$$

where $(-x_n \vee 0)^\uparrow = (-x_n)^\uparrow \vee 0$. By assumption, $x_n - c_n e \rightarrow y$. By Theorem 7.4.4,

$$\|(-x_n)^\uparrow - z(-y)\| \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

where $z(-y)(t) = -y(0)$, $0 \leq t \leq T$. By Lemma 7.5.3,

$$\|(-x_n)^\uparrow \vee 0 - z(-y) \vee 0\| \rightarrow 0 \quad \text{as } n \rightarrow \infty .$$

We obtain the desired convergence by adding, using the fact that the second term has a continuous limit.

(b) Apply the argument of Theorem 7.4.4 to show that, for all $\epsilon > 0$, there exists n_1 such that $-x_n(t) > -y(0) - \epsilon$ for all t , $0 \leq t \leq T$, and all $n \geq n_1$. Since $y(0) \leq 0$, $-x_n(t) > -\epsilon$ for all t , $0 \leq t \leq T$, and all $n \geq n_1$. Thus,

$$(-x_n + c_n e, (-x_n) \vee 0 + c_n e) \rightarrow (-y, -y)$$

in $D([0, T], \mathbb{R}^2)$ with the appropriate strong topology. Then, by Theorem 7.4.3,

$$(-x_n + c_n e, (-x_n) \vee 0 + c_n e, (-x_n \vee 0)^\uparrow + c_n e) \rightarrow (-y, -y, -y) \quad (5.3)$$

in $D([0, T], \mathbb{R}^3)$ with the appropriate strong topology. Then, by applying subtraction to the first and third terms in (5.3), we get

$$\begin{aligned} \phi(x_n) &\equiv x_n + (-x_n \vee 0)^\uparrow \\ &= [(-x_n \vee 0)^\uparrow + c_n e] - [-x_n + c_n e] \\ &\rightarrow -y + y = 0e \end{aligned} \quad (5.4)$$

as $n \rightarrow \infty$. ■

7.6. Inverse

We now consider the inverse map.. It is convenient to consider the inverse map on the subset D_u of x in $D \equiv D([0, \infty), \mathbb{R})$ that are unbounded above and satisfy $x(0) \geq 0$. For $x \in D_u$, let the inverse of x be

$$x^{-1}(t) = \inf\{s \geq 0 : x(s) > t\}, \quad t \geq 0. \quad (6.1)$$

As before, let D_0 be the subset of x in D with $x(0) \geq 0$, and let $D_\uparrow$ and $D_{\uparrow\uparrow}$ be the subsets of nondecreasing and strictly increasing functions in D_0 . Let $D_{u\uparrow} \equiv D_u \cap D_\uparrow$ and $D_{u\uparrow\uparrow} \equiv D_u \cap D_{\uparrow\uparrow}$. Clearly,

$$D_{\uparrow\uparrow} \subseteq D_\uparrow \subseteq D_u \subseteq D_0.$$

7.6.1. The M_1 Topology

Even for the M_1 topology, there are complications at the left endpoint of the domain $[0, \infty)$.

Example 7.6.1. *Complications at the left endpoint of the domain.* To see that the inverse map from $(D_{\uparrow}, U)$ to $(D_{\uparrow}, M_1)$ is in general not continuous, let $x(t) = 0$, $0 \leq t < 1$, and $x(t) = t$, $t \geq 1$; Let $x_n = t/n$, $0 \leq t < 1$ and $x_n(t) = t$, $t \geq 1$. Then $\|x_n - x\|_{\infty} = n^{-1} \rightarrow 0$, but $x_n^{-1}(0) = 0 \not\rightarrow 1 = x^{-1}(0)$, so that $x_n^{-1} \not\rightarrow x^{-1}$ (M_1). ■

To avoid the problem in Example 7.6.1, we can require that $x^{-1}(0) = 0$. To develop an equivalent condition, let $D_{\epsilon}^{\uparrow}$ be the subset of functions x in D_u such that $x(t) = 0$ for $0 \leq t \leq \epsilon$.

Then let

$$D_u^* \equiv \bigcap_{n=1}^{\infty} (D_{u,n^{-1}})^c. \quad (6.2)$$

Lemma 7.6.1. (measurability of D_u^*) *With the J_1 , M_1 or M_2 topology, D_u^* in (6.2) is a G_{δ} subset of D_u and*

$$D_u^* = \{x \in D_u : x^{-1}(0) = 0\}. \quad (6.3)$$

Let $D_{u\uparrow}^* \equiv D_{\uparrow} \cap D_u^*$. A key property of $D_{u\uparrow}^*$, not shared by $D_{u\uparrow}$ because of the complication at the origin, is that parametric representation (u, r) for x directly serve as parametric representations for x^{-1} when we switch the roles of the components u and r .

Lemma 7.6.2. (switching the roles of u and r) *For $x \in D_{u\uparrow}^*$, the graph Γ_x serves as the graph of $\Gamma_{x^{-1}}$ with the axes switched. Thus, $(u, r) \in \Pi(x)$ if and only if $(r, u) \in \Pi(x^{-1})$, where $\Pi(x)$ is the set of M_1 parametric representations.*

Corollary 7.6.1. (continuity on (D_u^*, M_1)) *The inverse map from (D_u^*, M_1) to $(D_{u\uparrow}, M_1)$ is continuous.*

Proof. First apply Theorem 7.4.1 for the supremum. Then apply Lemma 7.6.2. ■

We now generalize Corollary 7.6.1 by only requiring that the limit be in D_u^* .

Theorem 7.6.1. (measurability and continuity at limits in D_u^*) *The inverse map in (6.1) from (D_u, M_2) to $(D_{u\uparrow}, M_1)$ is measurable and continuous at $x \in D_u^*$, i.e., for which $x^{-1}(0) = 0$.*

Proof. First, recalling that the Borel σ -field on D coincides with the Kolmogorov σ -field generated by the projections, measurability follows from Lemma ??; it suffices to show that $\{x : x^{-1}(t) \leq a\}$ is measurable. However,

$$\begin{aligned} \{x : x^{-1}(t) \leq a\} &= \bigcap_{j=1}^{\infty} \bigcap_{k=1}^{\infty} \{x : x^{-1}((t + j^{-1})-) \leq a + k^{-1}\} \\ &= \bigcap_{j=1}^{\infty} \bigcap_{k=1}^{\infty} \{x : x^{\leftarrow}((t + j^{-1})) \leq a + k^{-1}\} \\ &= \bigcap_{j=1}^{\infty} \bigcap_{k=1}^{\infty} \{x : x(a + k^{-1}) \geq t + j^{-1}\}, \end{aligned} \quad (6.4)$$

which is measurable. Next we turn to continuity. For any $x \in D_u$, $x^{-1} = (x^{\uparrow})^{-1}$, so it suffices to start from $x_n^{\uparrow} \rightarrow x^{\uparrow}$. By Theorem 7.4.1, the assumed convergence $x_n \rightarrow x$ in (D_u, M_2) implies that $x_n^{\uparrow} \rightarrow x^{\uparrow}$ in $(D_{\uparrow}, M_2)$. However, the M_1 and M_2 topologies coincide in $D_{\uparrow}$. So $x_n^{\uparrow} \rightarrow x^{\uparrow}$ in $(D_{\uparrow}, M_1)$. Since $x \in D_u^*$, $x^{\uparrow} \in D_{u,\uparrow}^*$. However, we need not have $x_n^{\uparrow} \in D_{u,\uparrow}^*$. We could directly apply Lemma 7.6.2 if $x_n^{\uparrow} \in D_{\uparrow}^*$ for all sufficiently large n . Hence suppose that is not the case. Then there exists a subsequence $\{x_{n_k}^{\uparrow}\}$ with $x_{n_k}^{\uparrow} \notin D_{u,\uparrow}^*$ for all n_k . Necessarily, then, $x_{n_k}^{\uparrow}(0) = 0$ for all n_k . Since $x_n^{\uparrow} \rightarrow x^{\uparrow}$, we can conclude that $x^{\uparrow}(0) = 0$. Since $x^{\uparrow}$ is right continuous and $x^{\uparrow} \in D_{u,\uparrow}^*$, for any $\epsilon > 0$, there exists δ , $0 < \delta < \epsilon/2$, such that $\delta \in \text{Disc}(x^{\uparrow})^c$ and $0 < x^{\uparrow}(\delta) < \epsilon/2$. Let n_0 then be such that $|x_{n_k}^{\uparrow}(0) - x^{\uparrow}(0)| < \epsilon/2$ and $|x_{n_k}^{\uparrow}(\delta) - x^{\uparrow}(\delta)| < \epsilon/2$ for all $n \geq n_0$. Hence, for $n \geq n_0$, we can define an approximation to $x_{n_k}^{\uparrow}$ which belongs to $D_{u,\uparrow}^*$. In particular, let $x_{n_k}^*(0) = x_{n_k}^{\uparrow}(0) = 0$ and let $x_{n_k}^*(t) = x_{n_k}^{\uparrow}(t)$ for all $t \geq \delta$ and let $x_{n_k}^*$ be defined by linear interpolation in $[0, \delta]$. Then $x_{n_k}^* \in D_{u,\uparrow}^*$, $\|x_{n_k}^* - x_{n_k}^{\uparrow}\| < \epsilon$ and $\|(x_{n_k}^*)^{-1} - x_{n_k}^{\uparrow-1}\| < \epsilon$ for all $n_k \geq n_0$. For $n \geq n_0$ such that $x_n^{\uparrow} \in D_{u,\uparrow}^*$, let $x_n^* = x_n^{\uparrow}$. Since ϵ was arbitrary, we can choose x_n^* such that $x_n^* \rightarrow x^{\uparrow}$ (M_1), $\|x_n^* - x_n^{\uparrow}\| \rightarrow 0$ and $\|(x_n^*)^{-1} - x_n^{\uparrow-1}\| \rightarrow 0$ as $n \rightarrow \infty$. By Lemma 7.6.2, $(x_n^*)^{-1} \rightarrow x^{-1}$ (M_1). Since $\|(x_n^*)^{-1} - x_n^{\uparrow-1}\| \rightarrow 0$, $x_n^{\uparrow-1} \rightarrow x^{-1}$ (M_1) as well. ■

Corollary 7.6.2. . (continuity at strictly increasing functions) *The inverse map from (D_u, M_2) to $(D_{u,\uparrow}, U)$ is continuous at $x \in D_{u,\uparrow}$.*

Proof. First, $D_{u,\uparrow\uparrow} \subseteq D_{u,\uparrow}^*$, so that we can apply Theorem 7.6.1 to get $x_n^{-1} \rightarrow x^{-1}$ in $(D_{u,\uparrow}, M_1)$. However, by Lemma ??, $x^{-1} \in C$ when $x \in D_{u,\uparrow\uparrow}$. Hence the M_1 convergence $x_n^{-1} \rightarrow x^{-1}$ actually holds in the stronger topology of uniform convergence over compact subsets. ■

7.6.2. The M'_1 Topology

For cases in which the condition $x^{-1}(0) = 0$ in Theorem 7.6.1 is not satisfied, we can modify the M_1 and M_2 topologies to obtain convergence, following Puhalskii and Whitt (1997). With these new weaker topologies, which we call M'_1 and M'_2 , we do not require that $x_n(0) \rightarrow x(0)$ when $x_n \rightarrow x$. We construct the new topologies by extending the graph of each function x by appending the segment $[0, x(0)] \equiv \{\alpha 0 + (1 - \alpha)x(0) : 0 \leq \alpha \leq 1\}$. Let the new graph of $x \in D$ be

$$\begin{aligned} \Gamma'_x &= \{(z, t) \in \mathbb{R}^k \times [0, \infty) : z = \alpha x(t) + (1 - \alpha)x(t-) \\ &\quad \text{for } 0 \leq \alpha \leq 1 \text{ and } t \geq 0\}, \end{aligned} \quad (6.5)$$

where $x(0-) \equiv 0$. Let $\Pi'(x)$ and $\Pi'_2(x)$ be the sets of all M_1 and M_2 parametric representations of Γ'_x , defined just as before. We say that $x_n \rightarrow x$ in (D, M'_i) if there exist parametric representations $(u_n, r_n) \in \Pi'(x_n)$ and $(u, r) \in \Pi'(x)$, where Π' is the set of M'_1 and M'_2 parametric representations, such that

$$\|u_n - u\|_t \vee \|r_n - r\|_t \rightarrow 0 \quad \text{as } n \rightarrow \infty \quad \text{for each } t > 0. \quad (6.6)$$

With the M'_1 topologies, we obtain a cleaner statement than Lemma 7.6.2.

Lemma 7.6.3. (graphs of the inverse with the M'_1 topology) *For $x \in D_{u,\uparrow}$, the graph Γ'_x serves as the graph $\Gamma'_{x^{-1}}$ with the axes switched, so that $(u, r) \in \Pi'(x)$ ($\Pi'_2(x)$) if and only if $(r, u) \in \Pi'(x^{-1})$ ($\Pi'_2(x^{-1})$).*

Thus we get an alternative to Theorem 7.6.1.

Theorem 7.6.2. (continuity in the M'_1 topology) *The inverse map in (6.1) from (D_u, M'_2) to $(D_{u,\uparrow}, M'_1)$ is continuous.*

Proof. By the M'_2 analog of Theorem 7.4.1, if $x_n \rightarrow x$ in (D_u, M'_2) , then $x_n^\uparrow \rightarrow x^\uparrow$ in $(D_{u,\uparrow}, M'_2)$. Since the M'_2 topology coincides with the M'_i topology on $D_\uparrow$, we get $x_n^\uparrow \rightarrow x^\uparrow$ in $(D_{u,\uparrow}, M'_1)$. By Lemma 7.6.3, we get $(x_n^\uparrow)^{-1} \rightarrow (x^\uparrow)^{-1}$ in $(D_{u,\uparrow}, M'_1)$. That gives the desired result because $(x^\uparrow)^{-1} = x^{-1}$ for all $x \in D_u$. ■

An alternative approach to the difficulty at the origin besides M'_i topology on $D_u([0, \infty), \mathbb{R})$ is the ordinary M_i topology on $D_u((0, \infty), \mathbb{R})$. The difficulty at the origin goes away if we ignore it entirely, which we can do by making the function domain $(0, \infty)$ for the image of the inverse functions.

In particular, Theorem 7.6.2 implies the following corollary.

Corollary 7.6.3. (continuity when the origin is removed from the domain) *The inverse map in (6.1) from $D_u([0, \infty), M_2)$ to $D_{u,\uparrow}((0, \infty), M_1)$ is continuous.*

Proof. Since the M'_2 topology is weaker than M_2 , if $x_n \rightarrow x$ in $D_u([0, \infty), M_2)$, then $x_n \rightarrow x$ in $D_u([0, \infty), M'_2)$. Apply Theorem 7.6.2 to get $x_n^{-1} \rightarrow x^{-1}$ in $D_{u,\uparrow}([0, \infty), M'_1)$. That implies $x_n^{-1} \rightarrow x^{-1}$ for the restrictions in $D_\uparrow([t_1, t_2], M_1)$ for all $t_1, t_2 \in \text{Disc}(x^{-1})^c$, which in turn implies that $x_n^{-1} \rightarrow x^{-1}$ in $D_{u,\uparrow}((0, \infty), M_1)$. ■

However, in general we cannot work with the inverse on $D_u((0, \infty), \mathbb{R})$. We can obtain positive results if all the functions are required to be monotone. The following result is elementary.

Theorem 7.6.3. (equivalent characterizations of convergence for monotone functions) *For $x_n, n \geq 1, x \in D_{u,\uparrow}([0, \infty), \mathbb{R})$, the following are equivalent:*

$$x_n \rightarrow x \quad \text{in} \quad D_{u,\uparrow}((0, \infty), \mathbb{R}, M_1) ; \quad (6.7)$$

$$x_n \rightarrow x \quad \text{in} \quad D_{u,\uparrow}([0, \infty), \mathbb{R}, M'_1) ; \quad (6.8)$$

$$x_n(t) \rightarrow x(t) \quad \text{for all } t \text{ in a dense subset of } (0, \infty) ; \quad (6.9)$$

$$x_n^{-1} \rightarrow x^{-1} \quad \text{in} \quad D((0, \infty), \mathbb{R}, M_1) ; \quad (6.10)$$

$$x_n^{-1} \rightarrow x^{-1} \quad \text{in} \quad D([0, \infty), \mathbb{R}, M'_1) ; \quad (6.11)$$

$$x_n^{-1}(t) \rightarrow x^{-1}(t) \quad \text{for all } t \text{ in a dense subset of } (0, \infty). \quad (6.12)$$

Proof. Theorem 7.6.2 implies the equivalence of (6.8) and (6.11). Clearly, (6.8) $\rightarrow$ (6.7) $\rightarrow$ (6.9), so that (6.11) $\rightarrow$ (6.10) $\rightarrow$ (6.12). It thus suffices to show that (6.9) $\rightarrow$ (6.8). For any $\epsilon > 0$, we can find t and n_0 such that $0 < t < \epsilon$, $t \in \text{Disc}(x)$ and $|x_n(t) - x(t)| < \epsilon$ for $n \geq n_0$. Let $n_1 \geq n_0$ be such that $d_{M'_2}(x_n, x) < \epsilon$ for the restrictions to $[t, t']$ for any $t' > t$ with $t' \in \text{Disc}(x)^c$. Since x_n and x are nondecreasing and nonnegative, the bounds $d_{M'_2}(x_n, x) < \epsilon$ over $[t, t']$ and $|x_n(t) - x(t)| < \epsilon$ imply that $d_{M'_2}(x_n, x) < \epsilon$ for the restrictions over $[0, t']$. Since ϵ and t' were arbitrary, $x_n \rightarrow x$ in $D_\uparrow([0, \infty), \mathbb{R}, M'_2)$, but the M'_2 and M'_1 topologies are equivalent on $D_\uparrow$. ■

In general, convergence in $D([0, \infty), \mathbb{R}, M'_1)$ provides stronger control of the behavior at the origin than convergence in $D((0, \infty), \mathbb{R}, M_1)$. Nothing more is omitted from Section 13.6 of the book.

7.7. Inverse with Centering

We continue considering the inverse map, but now with centering. We start by considering linear centering. In particular, we consider when a limit for $c_n(x_n - e)$ implies a limit for $c_n(x_n^{-1} - e)$ when $x_n \in D_u \equiv D_u([0, \infty), \mathbb{R})$ and $c_n \rightarrow \infty$. By considering the behavior at one t , it is natural to anticipate that we should have $c_n(x_n^{-1} - e) \rightarrow -y$ when $c_n(x_n - e) \rightarrow y$. A first step for the M topologies is to apply Theorem 7.4.2, which yields limits for $c_n(x_n^\uparrow - e)$. Thus for the M topologies, it suffices to assume that $x_n \in D_\uparrow$.

Now we state the main limit theorem for inverse functions with centering.

Theorem 7.7.1. *Suppose that $c_n(x_n - e) \rightarrow y$ as $n \rightarrow \infty$ in $D([0, \infty), \mathbb{R})$ with one of the topologies M_2 , M_1 or J_1 , where $x_n \in D_u$, $c_n \rightarrow \infty$ and $y(0) = 0$.*

(a) *If the topology is M_2 or M_1 , then $c_n(x_n^{-1} - e) \rightarrow -y$ as $n \rightarrow \infty$ with the same topology.*

(b) *If the topology is J_1 and if y has no positive jumps, then $c_n(x_n^{-1} - e) \rightarrow -y$ as $n \rightarrow \infty$.*

Proof. (a) The proof is easy for the M_i topologies when $x_n \in D_u^*$ for all sufficiently large n . First, given $c_n(x_n - e) \rightarrow y$ (M_i), we can apply Theorem 7.4.2 (a) to conclude that $c_n(x_n^\uparrow - e) \rightarrow y$ (M_i). Hence we can assume that $x_n \in D_\uparrow^*$. Thus there exist parametric representations $(u_n, r_n) \in \Pi(c_n(x_n - e))$ and $(u, r) \in \Pi(x)$ of the appropriate type such that $\|u_n - u\|_t \vee \|r_n - r\|_t \rightarrow 0$ as $n \rightarrow \infty$ for all $t > 0$. Then $(u'_n, r_n) \in \Pi(x_n)$ for $u'_n = c_n^{-1}u_n + r_n$. Since $x_n \in D_\uparrow^*$ for n sufficiently large, $(r_n, u'_n) \in \Pi(x_n^{-1})$ and $(c_n(r_n - u'_n), u'_n) \in \Pi(c_n(x_n^{-1} - e))$ for sufficiently large n . However,

$$c_n(r_n - u'_n) = -u_n \quad (7.1)$$

and

$$\|u'_n - r\|_t \rightarrow 0 \quad \text{as } n \rightarrow \infty \text{ for all } t > 0, \quad (7.2)$$

so that $c_n(x_n^{-1} - e) \rightarrow -y$ (M_i) as $n \rightarrow \infty$. However, in general we need not have $x_n \in D_u^*$ for all sufficiently large n . So, suppose that we do not. We then only have $x_n \in D_u$ for all n . As before, we can apply Theorem 7.4.2 to show that it suffices to assume that $x_n \in D_\uparrow$ for all n . We now show that we can approximate $x_n \in D_\uparrow$ by $x_n^* \in D_\uparrow^*$ for all n sufficiently large, so that

$$c_n\|x_n - x_n^*\| \rightarrow 0 \quad \text{and} \quad c_n\|x_n^{-1} - (x_n^*)^{-1}\| \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (7.3)$$

The limits in (7.3) plus the triangle inequality imply that

$$d(c_n(x_n^* - e), y) \leq d(c_n(x_n - e), y) + c_n \|x_n - x_n^*\| \rightarrow 0 \quad \text{as } n \rightarrow \infty \quad (7.4)$$

and

$$\begin{aligned} d(c_n(x_n^{-1} - e), -y) \\ \leq \|c_n(x_n^{-1} - (x_n^*)^{-1})\| + d(c_n((x_n^*)^{-1} - e), -y) \rightarrow 0 \end{aligned} \quad (7.5)$$

as $n \rightarrow \infty$, where d is the M_i metric. Thus, the remaining problem is to construct $x_n^* \in D_\uparrow^*$ satisfying (7.3). Since $y(0) = 0$ and $y \in D$, for all $\epsilon > 0$, there exists δ_1 such that $v(y, 0, \delta_1) < \epsilon/2$. Since $c_n(x_n - e) \rightarrow y$ (M_2), there exists n_0 and δ_2 such that $v(c_n(x_n - e), y, 0, \delta/2) < \epsilon/2$ for all $n \geq n_0$. Thus

$$t - c_n^{-1}\epsilon < x_n(t) \leq t + c_n^{-1}\epsilon \quad (7.6)$$

for all $n \geq n_0$ and t with $0 \leq t \leq \delta \equiv \delta_1 \wedge \delta_2$. By Lemma ??,

$$t + c_n^{-1}\epsilon > x_n^{-1}(t-) \geq t - c_n^{-1}\epsilon \quad (7.7)$$

for all $n \geq n_0$ and t with $0 \leq t \leq \delta - c_n^{-1}\epsilon$. Now choose $n_1 \geq n_0$ so that $c_n^{-1}\epsilon < \delta/4$ for all $n \geq n_1$. Then, by (7.6), for $n \geq n_1$,

$$0 < x_n(\delta/4) < \delta/2 \quad (7.8)$$

and (7.7) holds for $0 \leq t \leq 3\delta/4$. Hence, if $n \geq n_1$ and $x_n \notin D_\uparrow^*$, we can construct $x_n^* \in D_\uparrow^*$ by letting $x_n^*(0) = x_n(0) = 0$, $x_n^*(t) = x_n(t)$, $t \geq \delta/4$, and letting x_n^* be defined by linear interpolation for t in $[0, \delta/4]$. By (7.8), $x_n^* \in D_\uparrow^*$. Since x_n^* is defined by linear interpolation over $[0, \delta/4]$, for $n \geq n_1$,

$$\|c_n(x_n^* - e)\|_{\delta/4} = \max\{c_n(x_n - e)(0), c_n(x_n - e)(\delta/4)\} \leq \epsilon, \quad (7.9)$$

so that

$$\|c_n(x_n^* - x_n)\| \leq \|c_n(x_n - e)\|_{\delta/4} + \|c_n(x_n^* - e)\|_{\delta/4} \leq 2\epsilon. \quad (7.10)$$

Similarly, $(x_n^*)^{-1}(t) = x_n^{-1}(t)$ for $t \leq x_n(\delta/4) < \delta/2$ and $n \geq n_1$, so that by (7.7)

$$\|c_n((x_n^*)^{-1} - x_n^{-1})\| \leq \|c_n(x_n^{-1} - e)\|_{\delta/2} + \|c_n((x_n^*)^{-1} - e)\|_{\delta/2} \leq 2\epsilon. \quad (7.11)$$

Since ϵ was arbitrary, (7.10) and (7.11) imply (7.3), as required.

(b) Since $c_n(x_n - e) \rightarrow y$ (J_1) and $c_n \rightarrow \infty$, $\|x_n - e\|_t \rightarrow 0$ as $n \rightarrow \infty$ for all $t > 0$. By Corollary 7.6.2, $\|x_n^{-1} - e\|_t \rightarrow 0$ as $n \rightarrow \infty$ for each

$t > 0$. By Theorem 7.2.2, we can apply the composition map to obtain $c_n(x_n \circ x_n^{-1} - x_n^{-1}) \rightarrow y(J_1)$. Hence it suffices to show that $c_n \|x_n \circ x_n^{-1} - e\|_t \rightarrow 0$ as $n \rightarrow \infty$ for all $t > 0$. However, by Corollary ??,

$$\begin{aligned} c_n \|x_n \circ x_n^{-1} - e\|_t &\leq c_n J_{x_n^{-1}(t)}(x_n) \\ &= J_{x_n^{-1}(t)}(c_n(x_n - e)), \end{aligned} \quad (7.12)$$

where $J_t(x)$ is the maximum jump of x over $[0, t]$, treating $x(0-)$ as 0. Since $c_n(x_n - e) \rightarrow y$, $y(0) = 0$ and y has no positive jumps, $J_t(c_n(x_n - e)) \rightarrow 0$ as $n \rightarrow \infty$ for all $t > 0$, which implies the desired conclusion. ■

Nothing else is omitted from Section 13.7 of the book.

7.8. Counting Functions

Inverse functions or first-passage-time functions are closely related to counting functions. A counting function is defined in terms of a sequence $\{s_n : n \geq 0\}$ of nondecreasing nonnegative real numbers with $s_0 = 0$. We can think of s_n as the partial sum

$$s_n \equiv x_1 + \cdots + x_n, \quad n \geq 1, \quad (8.1)$$

by simply writing $x_i \equiv s_i - s_{i-1}$, $i \geq 1$. The associated *counting function* $\{c(t) : t \geq 0\}$ is defined by

$$c(t) \equiv \max\{k \geq 0 : s_k \leq t\}, \quad t \geq 0. \quad (8.2)$$

To have $c(t)$ finite for all $t > 0$, we assume that $s_n \rightarrow \infty$ as $n \rightarrow \infty$. We can reconstruct the sequence $\{s_n\}$ from $\{c(t) : t \geq 0\}$ by

$$s_n = \inf\{t \geq 0 : c(t) \geq n\}, \quad n \geq 0. \quad (8.3)$$

The sequence $\{s_n\}$ and the associated function $\{c(t) : t \geq 0\}$ can serve as sample paths for a stochastic point process on the nonnegative real line. Then there are (countably) infinitely many points with the n^{th} point being located at s_n . The summands x_n are then the intervals between successive points. The most familiar case is when the sequence $\{x_n : n \geq 1\}$ constitutes the possible values from a sequence $\{X_n : n \geq 1\}$ of i.i.d. random variables with values in $\mathbb{R}_+$. Then the counting function $\{c(t) : t \geq 0\}$ constitutes a possible sample path of an associated renewal counting process $\{C(t) : t \geq 0\}$; see Section 7.3 of the book.

Paralleling Lemma 13.6.3 in the book, we have the following basic inverse relation for counting functions.

Lemma 7.8.1. *For any nonnegative integer n and nonnegative real number t ,*

$$s_n \leq t \quad \text{if and only if} \quad c(t) \geq n. \quad (8.4)$$

We can put counting functions in the setting of inverse functions on $D_\uparrow$ by letting

$$y(t) \equiv s_{[t]}, t \geq 0. \quad (8.5)$$

To have $y \in D_\uparrow$, we use the assumption that $s_n \rightarrow \infty$ as $n \rightarrow \infty$. if all the summands are strictly positive then

$$y^{-1}(t) = c(t) + 1, \quad t \geq 0, \quad (8.6)$$

where y^{-1} is the image of the inverse map in (6.1) applied to y in (8.5). With (8.6), limits for counting functions can be obtained by applying results in the previous two sections.

The connection to the inverse map can also be made when the summands x_i are only nonnegative. To do so, we observe that the counting function c is a time-transformation of y^{-1} . both are right-continuous, but $c(t) < y^{-1}(t)$. In particular, c and y can be expressed in terms of each other.

Lemma 7.8.2. (relation between counting functions and inverse functions)
For y in (8.5) and c in (8.2),

$$c(t) = y^{-1}(y(y^{-1}(t)-)-), \quad t \geq 0, \quad (8.7)$$

$$c(t) = y^{-1}(t-) \quad \text{for all} \quad t \in \text{Disc}(c) = \text{Disc}(y^{-1}), \quad (8.8)$$

$$y^{-1}(t) = c(c^{-1}(c(t))), \quad t \geq 0. \quad (8.9)$$

The three functions y , y^{-1} and c are depicted for a typical initial segment of a sequence $\{s_n : n \geq 0\}$ in Figure 13.1 of the book. We can apply (8.7)–(8.9) in Lemma 7.8.1 to show that limits for scaled counting functions with centering, are equivalent to limits for scaled inverse functions. We use the fact that the M topologies are not altered by changing to the left limits, because the graph is unchanged. We first consider the case of no centering; afterwards we consider the case of centering. When there is no centering, the M_1 and M_2 topologies coincide and reduce to pointwise convergence on a dense subset of $\mathbb{R}_+$ including 0.

Consider a sequence of counting functions $\{\{c_n(t) : t \geq 0\} : n \geq 1\}$ with associated processes

$$y_n^{-1}(t) \equiv c_n(c_n^{-1}(c_n(t))), \quad t \geq 0, \quad (8.10)$$

$y_n = (y_n^{-1})^{-1}$. Form scaled functions by setting

$$\mathbf{c}_n(t) = n^{-1}c_n(a_nt) \quad \text{and} \quad \mathbf{y}_n(t) = a_n^{-1}y_n(nt), \quad t \geq 0, \quad (8.11)$$

where a_n are positive real numbers with $a_n \rightarrow \infty$. Note that

$$\mathbf{c}_n^{-1}(t) = a_n^{-1}c_n^{-1}(nt) \quad \text{and} \quad \mathbf{y}_n^{-1}(t) = n^{-1}y_n(a_nt), \quad t \geq 0. \quad (8.12)$$

Theorem 7.8.1. (asymptotic equivalence of limits for scaled processes)
Suppose that $\mathbf{y}_n \in D_{u,\uparrow}$, $n \geq 1$, for $\mathbf{y}_n$ in (8.11). Then any one of the limits $\mathbf{y}_n \rightarrow y$, $\mathbf{y}_n^{-1} \rightarrow y^{-1}$, $\mathbf{c}_n \rightarrow y^{-1}$ or $\mathbf{c}_n^{-1} \rightarrow y^{-1}$ in $D_{\uparrow}([0, \infty), \mathbb{R})$ with the M_2 ($= M_1$) topology, for y_n^{-1} , $\mathbf{c}_n$ and $\mathbf{c}_n^{-1}$ in (8.11) and (8.12), implies the others.

Proof. The equivalence between $\mathbf{y}_n \rightarrow y$ and $\mathbf{y}_n^{-1} \rightarrow y^{-1}$, and between $\mathbf{c}_n \rightarrow y^{-1}$ and $\mathbf{c}_n^{-1} \rightarrow y$ follow from Theorem 7.6.1. We can relate the limits $\mathbf{c}_n \rightarrow y^{-1}$ and $\mathbf{y}_n \rightarrow y$ by applying (8.6), after modifying the summands $x_{n,i}$ in the sequences $\{s_{n,k} : k \geq 0\}$ to make them strictly positive. We can show that the limits are unaltered by adding suitably small positive values to the summands. Given $\epsilon > 0$ and $\{x_n : n \geq 1\}$, let

$$x'_n = x_n + \epsilon 2^{-n}, \quad n \geq 1, \quad (8.13)$$

and let $x'_n = x'_1 + \cdots + x'_n$, $n \geq 1$, and $c'(t) = \max\{k \geq 0 : s'_n \leq t\}$, $t \geq 0$. Then

$$s_n \leq s'_n \leq s_n + \epsilon, \quad n \geq 0, \quad (8.14)$$

and

$$c((t - \epsilon) \vee 0) \leq c'(t) \leq c(t), \quad t \geq 0. \quad (8.15)$$

The actual limits we want to consider involve a sequence of sequences $\{\{s_{n,k} : k \geq 0\}, n \geq 1\}$ with $s_{n,0} = 0$ for each n . Let $\{\{c_n(t) : t \geq 0\}\}$ be the associated sequence of counting functions. Let $x'_{n,k}$, $s'_{n,k}$, $n'_n(t)$, $\mathbf{s}'_n$ and $\mathbf{n}'_n$ be associated quantities defined by the modification in (8.7), i.e., by letting

$$x'_{n,k} \equiv x_{n,k} + \epsilon_n 2^{-k}, \quad k \geq 1. \quad (8.16)$$

Given that scaled processes are formed as in (8.11) and (8.12). It is elementary that

$$\|\mathbf{y}_n - \mathbf{y}'_n\|_{\infty} \leq \epsilon_n / a_n \rightarrow 0 \quad (8.17)$$

so that, for appropriate choice of ϵ_n , e.g., $\epsilon_n = \epsilon$, $\epsilon_n/a_n \rightarrow 0$. The bound in (8.15) enables us to conclude that $\mathbf{c}_n \rightarrow c$ (M_2) if and only if $\mathbf{c}'_n \rightarrow c$ (M_2) by applying Corollary 12.11.6 in the book. Hence it suffices to assume that the sequences $\{s_{n,k} : k \geq 0\}$ are strictly increasing, which implies that (8.6) holds. Then, after scaling as in (8.11) and (8.12),

$$\|\mathbf{y}_n^{-1} - \mathbf{c}_n\|_\infty \leq 1/n \rightarrow 0,$$

which completes the proof. ■

We now apply the results for inverse maps with centering in Section 7.7 to obtain limits for counting functions with centering. Consider a sequence of counting functions $\{\{c_n(t) : t \geq 0\} : n \geq 1\}$ associated with a sequence of nondecreasing sequences of nonnegative numbers $\{\{s_{n,k} : k \geq 0\} : n \geq 1\}$ defined as in (8.2). Let the scaled functions $\mathbf{c}_n$, $\mathbf{y}_n$, $\mathbf{c}_n^{-1}$ and $\mathbf{y}_n^{-1}$ be defined as in (8.10)–(8.12).

Theorem 7.8.2. (asymptotic equivalence of counting and inverse functions with centering) *Suppose that $\mathbf{y}_n \in D_\uparrow$, $n \geq 1$, $b_n \rightarrow \infty$ and $y(0) = 0$. Then any one of the limits $b_n(\mathbf{y}_n - e) \rightarrow y$, $b_n(\mathbf{c}_n - e) \rightarrow -y$, $b_n(\mathbf{y}_n^{-1} - e) \rightarrow -y$ or $b_n(\mathbf{c}_n^{-1} - e) \rightarrow y$ in $D([0, \infty), \mathbb{R})$ with the M_1 or M_2 topology, for $\mathbf{y}_n$, $\mathbf{c}_n$, and $\mathbf{y}_n^{-1}$ and $\mathbf{c}_n^{-1}$ in (8.11) and (8.12), implies the others with the same topology.*

Proof. The equivalence between $b_n(\mathbf{y}_n - e) \rightarrow y$ and $b_n(\mathbf{y}_n^{-1} - e) \rightarrow -y$ is contained in Theorem 7.7.1. Similarly, the equivalence between $b_n(\mathbf{c}_n - e) \rightarrow -y$ and $b_n(\mathbf{c}_n^{-1} - e) \rightarrow y$ is contained in Theorem 7.7.1. Let the topology be fixed at either M_1 or M_2 . Given $b_n(\mathbf{y}_n^{-1} - e) \rightarrow -y$, we have $\|\mathbf{y}_n^{-1} - e\|_t \rightarrow 0$ and $\|\mathbf{y}_n - e\|_t \rightarrow 0$ as $n \rightarrow \infty$ for each $t > 0$. For any $x \in D$, let $\hat{x}$ denote the associated left-limit function; i.e., $\hat{x}(t) = x(t-)$. Then $\mathbf{c}_n = \hat{\mathbf{y}}_n^{-1} \circ \hat{\mathbf{y}}_n \circ \mathbf{y}_n^{-1}$. Given $b_n(\mathbf{y}_n^{-1} - e) \rightarrow -y$, we have $\hat{\mathbf{y}}_n^{-1} \rightarrow e$, $\hat{\mathbf{y}}_n \rightarrow e$, $b_n(\hat{\mathbf{y}}_n^{-1} - e) \rightarrow -y$ and $b_n(\hat{\mathbf{y}}_n - e) \rightarrow y$, because the graphs are unchanged. Now we can apply the composition map to get $b_n(\hat{\mathbf{y}}_n^{-1} \circ \hat{\mathbf{y}}_n \circ \mathbf{y}_n^{-1} - \hat{\mathbf{y}}_n \circ \mathbf{y}_n^{-1}) \rightarrow -y$ and $b_n(\hat{\mathbf{y}}_n \circ \mathbf{y}_n^{-1} - \mathbf{y}_n^{-1}) \rightarrow y$. Hence, by Proposition ??, for each $t \in \text{Disc}(y)^c$, we have

$$\begin{aligned} b_n(\mathbf{c}_n - e)(t) &= b_n(\hat{\mathbf{y}}_n^{-1} \circ \hat{\mathbf{y}}_n \circ \mathbf{y}_n^{-1} - e)(t) \\ &= b_n(\hat{\mathbf{y}}_n^{-1} \circ \hat{\mathbf{y}}_n \circ \mathbf{y}_n^{-1} - \hat{\mathbf{y}}_n \circ \mathbf{y}_n^{-1})(t) \\ &\quad + b_n(\hat{\mathbf{y}}_n \circ \mathbf{y}_n^{-1} - \mathbf{y}_n^{-1})(t) + b_n(\mathbf{y}_n^{-1} - e)(t) \\ &\rightarrow -y(t) + y(t) - y(t) = -y(t). \end{aligned} \tag{8.18}$$

Now we apply Theorems 6.5.1 (iv) and 6.11.1 (iv). Let $w(x, \delta)$ be the M_i oscillation function over the interval $[0, t]$. By (8.8), the oscillations of $b_n(\mathbf{c}_n - e)$ coincide with the oscillations of $b_n(\mathbf{y}_n^{-1} - e)$ at discontinuity points of $\mathbf{c}_n$ and $\mathbf{y}_n^{-1}$. Moreover, in between such discontinuity points, they have identical maximum oscillations. Hence, for any interval $[0, t]$ with $t \in \text{Disc}(y)^c$,

$$w(b_n(\mathbf{c}_n - e), \delta) < w(b_n(\mathbf{y}_n^{-1} - e), \delta) . \quad (8.19)$$

Since $b_n(\mathbf{y}_n^{-1} - e) \rightarrow -y$ by assumption,

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w(b_n(\mathbf{y}_n^{-1} - e), \delta) = 0 \quad (8.20)$$

Consequently,

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w(b_n(\mathbf{c}_n - e), \delta) = 0 . \quad (8.21)$$

Hence, we can conclude that $b_n(\mathbf{c}_n - e) \rightarrow -y$.

To go the other way, suppose that $b_n(\mathbf{c}_n - e) \rightarrow -y$. Applying Theorem 7.7.1, we have $b_n(\mathbf{c}_n^{-1} - e) \rightarrow y$, $\mathbf{c}_n \rightarrow e$ and $\mathbf{c}_n^{-1} \rightarrow e$. Then, paralleling (8.18), we can apply (8.9) to obtain

$$\begin{aligned} b_n(\mathbf{y}_n^{-1} - e)(t) &= b_n(\mathbf{c}_n \circ \mathbf{c}_n^{-1} \circ \mathbf{c}_n - e)(t) \\ &= b_n(\mathbf{c}_n \circ \mathbf{c}_n^{-1} \circ \mathbf{c}_n - \mathbf{c}_n^{-1} \circ \mathbf{c}_n)(t) \\ &\quad + b_n(\mathbf{c}_n^{-1} \circ \mathbf{c}_n - \mathbf{c}_n)(t) + b_n(\mathbf{c}_n - e)(t) \\ &\rightarrow -y(t) + y(t) - y(t) = -y(t) \end{aligned} \quad (8.22)$$

for each $t \in \text{Disc}(y)^c$. Now let $w(x, \delta, t)$ denote the M_i oscillation function over the interval $[0, t]$ as a function of the right endpoint t . Then, paralleling (8.19), by (8.8), for all $t_1 \in \text{Disc}(y)^c$, there exists $t_2 > t_1$ with $t_2 \in \text{Disc}(y)^c$ such that

$$w(b_n(\mathbf{y}_n^{-1} - e), \delta, t_1) < w(b_n(\mathbf{c}_n - e), \delta, t_2) \quad (8.23)$$

for all n sufficiently large. Hence we can use the previous oscillation argument to conclude that $b_n(\mathbf{y}_n^{-1} - e) \rightarrow -y$. ■

7.9. Renewal-Reward Processes

Nothing was omitted from Section 13.9 in the book.

Chapter 8

Queueing Networks

8.1. Introduction

This chapter contains proofs omitted from Chapter 14 of the book, with the same title. Section 8.9 also contains supplementary material on the existence of a limiting stationary version for a general reflected process. With the exception of Section 8.9, the section and theorem numbering here parallels Chapter 14, so that the proofs should be easy to find.

Here is how this chapter is organized: We start in Section 8.2 by carefully defining the multidimensional reflection map and establishing its basic properties. Since the definition (Definition 8.2.1) is somewhat abstract, a key property is having the reflection map be well defined; i.e., we show that there exists a unique function satisfying the definition (Theorem 8.2.1). We also provide multiple characterizations of the reflection map, one alternative being as the unique fixed point of an appropriate operator (Theorem 8.2.2), while another is a basic complementarity property (Theorem 8.2.3).

A second key property of the multidimensional reflection map is Lipschitz continuity in the uniform norm on $D([0, T], \mathbb{R}^k)$ (Theorem 8.2.5). We also establish continuity of the multidimensional reflection map as a function of the reflection matrix, again in the uniform topology (Theorems 8.2.8 and 14.2.9 in the book). It is easy to see that the Lipschitz property is inherited when the metric on the domain and range is changed to d_{J_1} (Theorem 8.2.7). However, a corresponding direct extension for the SM_1 metric d_s does not hold. Much of the rest of the chapter is devoted to obtaining positive results for the M_1 topologies.

Section 8.3 provides yet another characterization of the multidimensional reflection map via an associated instantaneous reflection map on $\mathbb{R}^k$.

Sections 8.4 and 8.5 are devoted to obtaining the M_1 continuity results.

In Section 8.4 we establish properties of reflection of parametric representations. We are able to extend Lipschitz and continuity results from the uniform norm to the M_1 metrics when we can show that the reflection of a parametric representation can serve as the parametric representation of the reflected function. The results are somewhat complicated, because this property holds only under certain conditions.

In Sections 8.6 and 8.7, respectively, we apply the previous results to obtain heavy-traffic stochastic-process limits for stochastic fluid networks and conventional queueing networks. In the queueing networks we allow service interruptions. When there are heavy-tailed distributions or rare long service interruptions, the M_1 topologies play a critical role.

In Section 8.8 we consider the two-sided regulator and other reflection maps. The two-sided regulator is used to obtain heavy-traffic limits for single queues with finite waiting space, as considered in Section 2.3 and Chapter 5 of the book. With the scaling, the size of the waiting room is allowed to grow in the limit as the traffic intensity increases, but at a rate such that the limit process involves a two-sided regulator (reflection map) instead of the customary one-sided one. Like the one-sided reflection map, the two-sided regulator is continuous on (D^1, M_1) . Moreover, the content portion of the two-sided regulator is Lipschitz, but the two regulator portions (corresponding to the two barriers) are only continuous; they are not Lipschitz.

We also give general conditions for other reflection maps to have M_1 continuity and Lipschitz properties. For these, we require that the limit function to be reflected belong to D_1 , the subset of functions with discontinuities in only one coordinate at a time.

In Section 8.9 we show that reflected stochastic processes have proper limiting stationary distributions and proper limiting stationary versions (stochastic-process limits for the entire time-shifted processes) under very general conditions. Our main result, Theorem 8.9.1, establishes such limits for stationary ergodic net-input stochastic processes satisfying a natural drift condition (9.7). It is noteworthy that a proper limit can exist even if there is positive drift in some (but not all) coordinates. Theorem 8.9.1 is limited by having a special initial condition: starting out empty. Much of the rest of Section 8.9 is devoted to obtaining corresponding results for other initial conditions. Theorem 8.9.6 establishes convergence for all proper initial contents when the net input process is also a Lévy process with mutually independent coordinate processes. Theorem 8.9.6 covers limit processes obtained in the heavy-traffic limits for the stochastic fluid networks in Section 14.6 of the book.

8.2. The Multidimensional Reflection Map

We start by giving basic definitions and establishing alternative characterizations. Then we establish continuity and Lipschitz properties.

8.2.1. Definition and Characterization

Let $\mathcal{Q}$ be the set of all reflection matrices, i.e., the set of all column-stochastic matrices Q (with $Q_{i,j}^t \geq 0$ and $\sum_{j=1}^k Q_{i,j}^t \leq 1$) such that $Q^n \rightarrow 0$ as $n \rightarrow \infty$, where Q^n is the n^{th} power of Q .

Definition 8.2.1. (reflection map) *For any $x \in D^k \equiv D([0, T], \mathbb{R}^k)$ and any reflection matrix $Q \in \mathcal{Q}$, let the feasible regulator set be*

$$\Psi(x) \equiv \{w \in D_{\uparrow}^k : x + (I - Q)w \geq 0\} \quad (2.1)$$

and let the reflection map be $R \equiv (\psi, \phi) : D^k \rightarrow D^{2k}$ with regulator component

$$y \equiv \psi(x) \equiv \inf \Psi(x) \equiv \inf\{w : w \in \Psi(x)\} , \quad (2.2)$$

i.e.,

$$y^i(t) \equiv \inf\{w^i(t) \in \mathbb{R} : w \in \Psi(x)\} \quad \text{for all } i \text{ and } t , \quad (2.3)$$

and content component

$$z \equiv \phi(x) \equiv x + (I - Q)y . \quad (2.4)$$

It remains to show that the reflection map is well defined by Definition 8.2.1; i.e., we need to know that the feasible regulator set $\Psi(x)$ is nonempty and that its infimum y (which necessarily is well defined and unique for nonempty $\Psi(x)$) is itself an element of $\Psi(x)$, so that $z \in D^k$ and $z \geq 0$.

To show that $\Psi(x)$ in (2.1) is nonempty, we exploit the well known fact that the matrix $I - Q$ has nonnegative inverse.

Lemma 8.2.1. (nonnegative inverse of reflection matrix) *For all $Q \in \mathcal{Q}$, $I - Q$ is nonsingular with nonnegative inverse*

$$(I - Q)^{-1} = \sum_{n=0}^{\infty} Q^n ,$$

where $Q^0 = I$.

Proof. Note that

$$(I - Q)(I + Q + \cdots + Q^{n-1}) = I - Q^n. \quad (2.5)$$

Since $Q^n \rightarrow 0$ as $n \rightarrow \infty$, $I - Q^n \rightarrow I$ as $n \rightarrow \infty$, where I has determinant 1. Hence, for all sufficiently large n , the left and right sides of (2.5) have nonzero determinant. Since the determinant of the product of two matrices is the product of the determinants, the determinant of $I - Q$ must be nonzero, so that $I - Q$ must be nonsingular. Now multiply both sides of (2.5) by this inverse, which we have shown exists, to obtain

$$I + Q + \cdots + Q^{n-1} = (I - Q)^{-1}(I - Q^n).$$

Since the right side tends to the proper limit $(I - Q)^{-1}$ as $n \rightarrow \infty$, so does the left. ■

The key to showing that the infimum belongs to the feasibility set is a basic result about semicontinuous functions. Recall that a real-valued function x on $[0, T]$ is *upper semicontinuous* at a point t in its domain if

$$\limsup_{t_n \rightarrow t} x(t_n) \leq x(t)$$

for any sequence $\{t_n\}$ with $t_n \in [0, T]$ and $t_n \rightarrow t$ as $n \rightarrow \infty$. The function x is upper semicontinuous if it is upper semicontinuous at all arguments t in its domain.

Lemma 8.2.2. (preservation of upper semicontinuity) *Suppose that $\{x_s : s \in S\}$ is a set of upper semicontinuous real-valued function on a subinterval of $\mathbb{R}$. Then the infimum $\underline{x} \equiv \inf\{x_s : s \in S\}$ is also upper semicontinuous.*

Proof. For any t and $\epsilon > 0$ given, we need to find δ such that $\underline{x}(t') \leq \underline{x}(t) + \epsilon$ whenever $|t' - t| < \delta$. Since $\underline{x}$ is the infimum, for any t and ϵ , we can find $x \in \{x_s : s \in S\}$ such that $x(t) \leq \underline{x}(t) + \epsilon/2$. Since x is upper semicontinuous, there exists δ such that $x(t') \leq x(t) + \epsilon/2$ for all t' with $|t - t'| < \delta$. As a consequence,

$$\underline{x}(t') \leq x(t') \leq x(t) + \epsilon/2 \leq \underline{x}(t) + \epsilon$$

whenever $|t - t'| < \delta$. ■

Recall that $x^\uparrow \equiv \sup_{0 \leq s \leq t} x(s)$, $t \geq 0$, for $x \in D^1$. For $x \equiv (x^1, \dots, x^k) \in D^k$, let $x^\uparrow \equiv ((x^1)^\uparrow, \dots, (x^k)^\uparrow)$.

Theorem 8.2.1. (existence of the reflection map) *For any $x \in D^k$ and $Q \in \mathcal{Q}$,*

$$(I - Q)^{-1}[(-x)^\uparrow \vee 0] \in \Psi(x) , \quad (2.6)$$

so that $\Psi(x) \neq \emptyset$,

$$y \equiv \psi(x) \in \Psi(x) \subseteq D_\uparrow^k \quad (2.7)$$

for y in (2.2) and

$$z \equiv \phi(x) = x + (I - Q)y \geq 0 . \quad (2.8)$$

Proof. The proof is in the book. ■

We now characterize the regulator function $y \equiv \psi(x)$ as the unique fixed point of a mapping $\pi \equiv \pi_{x,Q} : D_\uparrow^k \rightarrow D_\uparrow^k$, defined by

$$\pi(w) = (Qw - x)^\uparrow \vee 0 \quad (2.9)$$

for $w \in D_\uparrow^k$. For this purpose, we use two elementary lemmas.

Lemma 8.2.3. (feasible regulator set characterization) *The feasible regulator set $\Psi(x)$ in (2.1) can be characterized by*

$$\Psi(x) = \{w \in D_\uparrow^k : w \geq \pi(w)\}$$

for π in (2.9).

Proof. The proof is in the book. ■

Lemma 8.2.4. (closed subset of D) *With the uniform topology on D , The feasible regulator set $\Psi(x)$ is a closed subset of $D_\uparrow^k$, while $D_\uparrow^k$ is a closed subset of D .*

Theorem 8.2.2. (fixed-point characterization) *For each $Q \in \mathcal{Q}$, the regulator map $y \equiv \psi(x) \equiv \psi_Q(x) : D^k \rightarrow D_\uparrow^k$ can be characterized as the unique fixed point of the map $\pi \equiv \pi_{x,Q} : D_\uparrow^k \rightarrow D_\uparrow^k$ defined in (2.9).*

Proof. The proof is in the book. ■

Theorem 8.2.3. (complementarity characterization) *A function y in the feasible regulator set $\Psi(x)$ in (2.1) is the infimum $\psi(x)$ in (2.2) if and only if the pair (y, z) for $z \equiv x + (I - Q)y$ satisfies the complementarity property*

$$\int_0^\infty z^i dy^i = 0, \quad 1 \leq i \leq k . \quad (2.10)$$

Proof. The proof is in the book. ■

8.2.2. Continuity and Lipschitz Properties

We now establish continuity and Lipschitz properties of the reflection map as a function of the function x and the reflection matrix Q . We use the *matrix norm*, defined for any $k \times k$ real matrix A by

$$\|A\| \equiv \max_j \sum_{i=1}^k |A_{i,j}| . \quad (2.11)$$

We use the maximum column sum in (2.11) because we intend to work with the column-substochastic matrices in $\mathcal{Q}$. Note that

$$\|A_1 A_2\| \leq \|A_1\| \cdot \|A_2\|$$

for any two $k \times k$ real matrices A_1 and A_2 . Also, using the sum (or l_1) norm

$$\|u\| \equiv \sum_{i=1}^k |u^i| \quad (2.12)$$

on $\mathbb{R}^k$, we have

$$\|Au\| \leq \|A\| \cdot \|u\| \quad (2.13)$$

for each $k \times k$ real matrix A and $u \in \mathbb{R}^k$. Indeed, we can also define the matrix norm by

$$\|A\| \equiv \max\{\|Au\| : u \in \mathbb{R}^n, \|u\| = 1\} , \quad (2.14)$$

using the sum norm in (2.12) in both places on the right. Then (2.11) becomes a consequence. Consistent with (2.12), we let

$$\|x\| \equiv \sup_{0 \leq t \leq T} \|x(t)\| \equiv \sup_{0 \leq t \leq T} \sum_{i=1}^k \|x^i(t)\| \quad (2.15)$$

for $x \in D([0, T], \mathbb{R}^k)$. Combining (2.13) and (2.15), we have

$$\|Ax\| \leq \|A\| \cdot \|x\| \quad (2.16)$$

for each $k \times k$ real matrix A and $x \in D([0, T], \mathbb{R}^k)$.

We use the following basic lemma.

Lemma 8.2.5. (reflection matrix norms) *For any $k \times k$ matrix $Q \in \mathcal{Q}$,*

$$\|Q\| \leq 1, \quad \|Q^k\| = \gamma < 1 \quad (2.17)$$

and

$$\|(I - Q)^{-1}\| \leq \frac{k}{1 - \gamma} . \quad (2.18)$$

Proof. The first relation in (2.17) is immediate. Since $Q^n \rightarrow 0$ for all Q in $\mathcal{Q}$, the Markov chain associated with Q^t is transient. Since the Markov chain has k states,

$$\sum_{j=1}^k (Q_{i,j}^t)^k < 1, \quad (2.19)$$

which, with (2.11), implies the second relation in (2.17). Probabilistically, if the probability of eventually exiting the state space $\{1, \dots, k\}$ of the Markov chain is 1, then the probability of immediately exiting the state space from some state must be positive. Then the probability of reaching that state or the exterior (leaving the state space) in one step must be positive from some other state. Proceeding on by induction, the state space must be exhausted after k steps, so that (2.19) holds. Finally,

$$\left\| \sum_{n=0}^{\infty} Q^n \right\| \leq \sum_{n=0}^{\infty} \|Q^n\| \leq \sum_{n=0}^{k-1} \|Q^n\| + \gamma \sum_{n=0}^{\infty} \|Q^n\|$$

so that (2.18) holds. ■

We now show that $\pi \equiv \pi_{x,Q}$ in (2.9) is a k -stage contraction map on $D_{\uparrow}^k$. Recall that for $x \in D$, $|x|$ denotes the function $\{|x(t)| : t \geq 0\}$ in D , where $|x(t)| = (|x^1(t)|, \dots, |x^k(t)|) \in \mathbb{R}^k$. Thus, for $x \in D$, $|x|^{\uparrow} = (|x^1|^{\uparrow}, \dots, |x^k|^{\uparrow})$, where $|x^i|^{\uparrow}(t) = \sup_{0 \leq s \leq t} |x^i(s)|$, $0 \leq t \leq T$.

Lemma 8.2.6. (π is a k -stage contraction) *For any $Q \in \mathcal{Q}$ and $w_1, w_2 \in D_{\uparrow}^k$,*

$$|\pi^n(w_1) - \pi^n(w_2)|^{\uparrow} \leq |Q^n(|w_1 - w_2|^{\uparrow})| \quad \text{for } n \geq 1, \quad (2.20)$$

so that

$$\|\pi^n(w_1) - \pi^n(w_2)\| \leq \|Q^n\| \cdot \|w_1 - w_2\| \leq \|w_1 - w_2\| \quad (2.21)$$

for $n \geq 1$ and

$$\|\pi^n(w_1) - \pi^n(w_2)\| \leq \gamma \|w_1 - w_2\| \quad \text{for } n \geq k,$$

where

$$\|Q^k\| \equiv \gamma < 1.$$

Hence

$$\|\pi^n(w) - \psi(x)\| \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Proof. The proof is in the book. ■

We now establish inequalities that imply that the reflection map is a Lipschitz continuous map on $(D, \|\cdot\|)$. We will use the stronger inequalities themselves in Section 8.9.

Theorem 8.2.4. (one-sided bounds) *For any $Q \in \mathcal{Q}$ and $x_1, x_2 \in D$,*

$$-(I - Q)^{-1}\eta_1(x_1 - x_2) \leq \psi(x_1) - \psi(x_2) \leq (I - Q)^{-1}\eta_1(x_2 - x_1) \quad (2.22)$$

where $\eta_1(x) \equiv (\hat{\eta}_1(x^1), \dots, \hat{\eta}_1(x^k))$ with $\hat{\eta}_1 : D^1 \rightarrow D^1$ defined by

$$\hat{\eta}_1(x^i) \equiv (x^i)^\uparrow \vee 0 .$$

Proof. The proof is in the book. ■

As a direct consequence of Theorem 8.2.4, we obtain the desired Lipschitz property.

Theorem 8.2.5. (Lipschitz property with uniform norm) *For any $Q \in \mathcal{Q}$ and $x_1, x_2 \in D$,*

$$\begin{aligned} \|\psi(x_1) - \psi(x_2)\| &\leq \|(I - Q)^{-1}\| \cdot \|x_1 - x_2\| \\ &\leq \sum_{n=0}^{\infty} \|Q^n\| \cdot \|x_1 - x_2\| \\ &\leq \frac{k}{1 - \gamma} \|x_1 - x_2\| , \end{aligned} \quad (2.23)$$

where $\gamma \equiv \|Q^k\| < 1$, and

$$\begin{aligned} \|\phi(x_1) - \phi(x_2)\| &\leq (1 + \|I - Q\| \cdot \|(I - Q)^{-1}\|) \|x_1 - x_2\| \\ &\leq \left(1 + \frac{2k}{1 - \gamma}\right) \|x_1 - x_2\| . \end{aligned} \quad (2.24)$$

Proof. The proof is in the book. ■

We now summarize some elementary but important properties of the reflection map.

Theorem 8.2.6. (reflection map properties) *The reflection map satisfies the following properties:*

- (i) adaptedness: *For any $x \in D$ and $t \in [0, T]$, $R(x)(t)$ depends upon x only via $\{x(s) : 0 \leq s \leq t\}$.*
- (ii) monotonicity: *If $x_1 \leq x_2$ in D , then $\psi(x_1) \geq \psi(x_2)$.*

(iii) rescaling: For each $x \in D([0, T], \mathbb{R}^k)$, $\eta \in \mathbb{R}^k$, $\beta > 0$ and γ nondecreasing right-continuous function mapping $[0, T_1]$ into $[0, T]$, $\eta + \beta(x \circ \gamma) \in D([0, T_1], \mathbb{R}^k)$ and

$$R(\eta + \beta(x \circ \gamma)) = \beta R(\beta^{-1}\eta + x) \circ \gamma .$$

(iv) shift: For all $x \in D$ and $0 < t_1 < t_2 < T$,

$$\psi(x)(t_2) = \psi(x)(t_1) + \psi(\phi(x)(t_1) + x(t_1 + \cdot) - x(t_1))(t_2 - t_1)$$

and

$$\phi(x)(t_2) = \phi(\phi(x)(t_1) + x(t_1 + \cdot) - x(t_1))(t_2 - t_1)$$

(v) continuity preservation: If $x \in C$, then $R(x) \in C$.

We can apply Theorems 8.2.5 and 8.2.6 (iii) to deduce that the reflection map inherits the Lipschitz property on (D, J_1) from (D, U) . Unfortunately, we will have to work harder to obtain related results for the M_1 topologies.

Theorem 8.2.7. (Lipschitz property with d_{J_1}) For any $Q \in \mathcal{Q}$, there exist constants K_1 and K_2 (the same as in Theorem 8.2.5) such that

$$d_{J_1}(\psi(x_1), \psi(x_2)) \leq K_1 d_{J_1}(x_1, x_2) \quad (2.25)$$

and

$$d_{J_1}(\phi(x_1), \phi(x_2)) \leq K_2 d_{J_1}(x_1, x_2) \quad (2.26)$$

for all $x_1, x_2 \in D$.

Proof. The proof is in the book. ■

We now want to consider the reflection map R as a function of the reflection matrix Q as well as the net input function x . We first consider the maps $\pi \equiv \pi_{x, Q}^n(0)$ in (2.9) and $\psi \equiv \psi_Q$ in (2.2) as functions of Q when Q is a strict contraction in the matrix norm (2.11), i.e., when $\|Q\| < 1$.

Theorem 8.2.8. (stability bounds for different reflection matrices) Let $Q_1, Q_2 \in \mathcal{Q}$ with $\|Q_1\| = \gamma_1 < 1$ and $\|Q_2\| = \gamma_2 < 1$. For all $n \geq 1$,

$$\|\pi_{x, Q_j}^n(0)\| \leq (1 + \gamma_j + \cdots + \gamma_j^{n-1})\|x\| \quad (2.27)$$

and

$$\|\pi_{x, Q_1}^n(0) - \pi_{x, Q_2}^n(0)\| \leq (1 + \gamma_2 + \cdots + \gamma_2^{n-1}) \frac{\|x\| \cdot \|Q_1 - Q_2\|}{1 - \gamma_1} , \quad (2.28)$$

so that

$$\|\psi_{Q_j}(x)\| \leq \frac{\|x\|}{1 - \gamma_j} \quad (2.29)$$

and

$$\|\psi_{Q_1}(x) - \psi_{Q_2}(x)\| \leq \frac{\|x\| \cdot \|Q_1 - Q_2\|}{(1 - \gamma_1)(1 - \gamma_2)}. \quad (2.30)$$

Proof. First

$$\|\pi_{x,Q_j}^1(0)\| = \|(-x)^\uparrow \vee 0\| \leq \|x\|.$$

Next, by induction,

$$\begin{aligned} \|\pi_{x,Q_j}^{n+1}(0)\| &= \|(Q_j \pi_{x,Q_j}^n(0) - x)^\uparrow \vee 0\| \\ &\leq \|Q_j\| \cdot \|\pi_{x,Q_j}^n(0)\| + \|x\| \\ &\leq \gamma_j(1 + \gamma_j + \cdots + \gamma_j^{n-1})\|x\| + \|x\| \\ &\leq (1 + \gamma_j + \cdots + \gamma_j^n)\|x\|. \end{aligned}$$

Similarly, by induction

$$\begin{aligned} \|\pi_{x,Q_1}^{n+1}(0) - \pi_{x,Q_2}^{n+1}(0)\| &\leq \|Q_1 \pi_{x,Q_1}^n(0) - Q_2 \pi_{x,Q_2}^n(0)\| \\ &\leq \|Q_1 \pi_{x,Q_1}^n(0) - Q_2 \pi_{x,Q_1}^n(0)\| + \|Q_2 \pi_{x,Q_1}^n(0) - Q_2 \pi_{x,Q_2}^n(0)\| \\ &\leq \|Q_1 - Q_2\| \cdot \|x\|/(1 - \gamma_1) + \|Q_2\| \cdot \|\pi_{x,Q_1}^n(0) - \pi_{x,Q_2}^n(0)\| \\ &\leq (1 + \gamma_2 + \cdots + \gamma_2^n)\|Q_1 - Q_2\| \cdot \|x\|/(1 - \gamma_1). \end{aligned}$$

Finally, since $\|\pi_{x,Q}^n(0) - \psi_Q(x)\| \rightarrow 0$ as $n \rightarrow \infty$, the final two bounds (2.29) and (2.30) follow. ■

Nothing more is omitted from Section 14.2 of the book.

8.3. The Instantaneous Reflection Map

Nothing has been deleted from this section in the book.

8.4. Reflections of Parametric Representations

In order to establish continuity and stronger Lipschitz properties of the reflection map R on D with the M_1 topologies, we would like to have $(R(u), r)$ be a parametric representation of $R(x)$ when (u, r) is a parametric representation of x . That is not always true, but we now obtain positive results in that direction.

Theorem 8.4.1. (reflections of parametric representations) *Suppose that $x \in D$, $(u, r) \in \Pi_s(x)$ and $r^{-1}(t) = [s_-(t), s_+(t)]$.*

(a) *If $t \in \text{Disc}(x)^c$, then*

$$R(u)(s) = R(x)(t) \quad \text{for } s_-(t) \leq s \leq s_+(t) .$$

(b) *If $t \in \text{Disc}(x)$, then*

$$R(u)(s_-(t)) = R(x)(t-) \quad \text{and} \quad R(u)(s_+(t)) = R(x)(t) .$$

(c) *If $t \in \text{Disc}(x)$ and $x(t) \geq x(t-)$, then*

$$\phi(u)(s) = \phi(x)(t-) + \left(\frac{u^j(s) - u^j(s_-(t))}{u^j(s_+(t)) - u^j(s_-(t))} \right) [x(t) - x(t-)]$$

for any j , $1 \leq j \leq k$, and

$$\psi(u)(s) = \psi(x)(t-) = \psi(x)(t) \quad \text{for } s_-(t) \leq s \leq s_+(t) ,$$

so that

$$R(u)(s) \in [R(x)(t-), R(x)(t)] \quad \text{for } s_-(t) \leq s \leq s_+(t) .$$

(d) *If $t \in \text{Disc}(x)$ and $x(t) \leq x(t-)$, then $\phi^i(u)$ and $\psi^i(u)$ are monotone in $[s_-(t), s_+(t)]$ for each i , so that*

$$R(u)(s) \in [[R(x)(t-), R(x)(t)]] \quad \text{for } s_-(t) \leq s \leq s_+(t) .$$

We can draw the desired conclusion that $(R(u), r)$ is a parametric representation of $R(x)$ if we can apply parts (c) and (d) of Theorem 8.4.1 to all jumps. Recall that D_+ (D_s) is the subset of D for which condition (c) (condition (c) or (d)) holds at all discontinuity points of x . For $x \in D_s$, the direction of the inequality is allowed to depend upon t .

Theorem 8.4.2. (preservation of parametric representations under reflection) *Suppose that $x \in D$ and $(u, r) \in \Pi_s(x)$.*

(a) *If $x \in D_+$, then $(R(u), r) \in \Pi_s(R(x))$.*

(b) *If $x \in D_s$, then $(R(u), r) \in \Pi_w(R(x))$.*

We also have an analog of Theorems 8.4.1 and 8.4.2 for the case $x \in D_s$ and $(u, r) \in \Pi_w(x)$.

Theorem 8.4.3. (preservation of weak parametric representations) *If $x \in D_s$ and $(u, r) \in \Pi_w(x)$, then $(R(u), r) \in \Pi_w(R(x))$.*

As a basis for proving Theorem 8.4.1, we exploit piecewise-constant approximations.

Lemma 8.4.1. (left and right limits) *For any $x \in D_c$, $(u, r) \in \Pi_s(x)$ and $r^{-1}(t) = [s_-(t), s_+(t)]$,*

$$R(u)(s_-(t)) = R(x)(t-) \quad \text{and} \quad R(u)(s_+(t)) = R(x)(t) . \quad (4.1)$$

In order to prove Lemma 8.4.1, we establish several other lemmas. First, the following property of the reflection map applied to a single jump at time t is an easy consequence of the definition of the reflection map. We consider the reflection map applied to the jump in two parts. Given the linear relationship in (2.4), it suffices to focus on only one of ψ or ϕ .

Lemma 8.4.2. (the case of a single jump) *For any $b_1, b_2 \in \mathbb{R}^k$, $0 < \beta < 1$ and $0 < t \leq T$,*

$$\phi(b_1 + b_2 I_{[t, T]})(u) = \phi(\phi(b_1 + \beta b_2 I_{[t, T]})(t) + (1 - \beta) b_2 I_{[t, T]})(u) \quad \text{for } t \leq u \leq T .$$

Lemma 8.4.3. (generalization) *For any $b_1, b_2 \in \mathbb{R}^k$ and right-continuous nondecreasing nonnegative real-valued function α on $[0, T]$ with $\alpha(0) = 0$,*

$$\phi(b_1 + \alpha b_2)(t) = \phi(b_1 + \alpha(t) b_2 I_{[0, T]})(t), \quad 0 \leq t \leq T . \quad (4.2)$$

Proof. Represent α as the uniform limit of nondecreasing nonnegative functions α_n in D_c . Then $\|\phi(b_1 + \alpha_n b_2) - \phi(b_1 + \alpha b_2)\| \rightarrow 0$ as $n \rightarrow \infty$ by the known continuity of ϕ in the uniform metric. Hence it suffices to assume that $\alpha \in D_c$. We then establish (4.2) by recursively considering the successive discontinuity points of α , using Lemma 8.4.2 and Theorem 8.2.6(iv). ■

Proof of Lemma 8.4.1. Any $x \in D_c$ can be represented as

$$x = \sum_{j=0}^m b_j I_{[t_j, T]}$$

for $0 = t_0 < t_1 < \dots < t_m \leq T$ and $b_j \in \mathbb{R}^k$ for $0 \leq j \leq m$. Thus t_j is the j^{th} discontinuity point of x . Let $[s_-(t_j), s_+(t_j)] = r^{-1}(t_j)$ for each j . Since $(u, r) \in \Pi_s(x)$ instead of just $\Pi_w(x)$, u can be expressed as

$$u = \sum_{j=0}^m \alpha_j b_j ,$$

where $\alpha_0(s) = 1$ for all s and, for $j \geq 1$, $\alpha_j : [0, 1] \rightarrow [0, 1]$ is continuous and nondecreasing with $\alpha_j(s) = 0$, $s \leq s_-(t_j)$ and $\alpha_j(s) = 1$, $s \geq s_+(t_j)$. We can now consider successive intervals $[s_-(t_j), s_+(t_j)]$ recursively exploiting Lemma 8.4.3. First, for any s with $0 \leq s \leq s_-(t_1)$.

$$\phi(u)(s) = \phi(b_0 I_{[0,1]})(s) = \phi(x)(0) = \phi_0(x(0)) .$$

Now assume that (4.1) holds for all $j \leq m-1$ and consider $s \in [s_-(t_m), s_+(t_m)]$. By the induction hypothesis, Lemma 8.4.3 and Theorem 8.2.6(iv),

$$\begin{aligned} \phi(u)(s) &= \phi(\phi(x)(t_{m-1}) + \alpha_m b_m I_{[s_-(t_m), 1]})(s) \\ &= \phi(\phi(x)(t_{m-1}) + \alpha_m(s) b_m I_{[s_-(t_m), 1]})(s) , \end{aligned}$$

so that (4.1) holds for t_m . ■

Proof of Theorem 8.4.1. (a) Since $t \in \text{Disc}(x)^c$, $u(s) = x(t)$ for $s_-(t) \leq s \leq s_+(t)$. Given $x \in D$ with $t \in \text{Disc}(x)^c$, it is possible to choose $x_n \in D_c$ such that $t \in \text{Disc}(x_n)^c$ for all n and $\|x_n - x\| \rightarrow 0$, by a slight strengthening of Theorem 6.2.2 in Section 6.2. By characterization (i) of M_1 convergence in Theorem 6.1 in Section V.6, given $(u, r) \in \Pi_s(x)$, we can find $(u_n, r_n) \in \Pi_s(x_n)$ such that

$$\|u_n - u\| \vee \|r_n - r\| \rightarrow 0 \quad \text{as } n \rightarrow \infty .$$

Since R is continuous in the uniform topology, $\|R(u_n) - R(u)\| \rightarrow 0$ and $\|R(x_n) - R(x)\| \rightarrow 0$ as $n \rightarrow \infty$. Let s_n be such that $r_n(s_n) = t$. Since $x_n \in D_c$ and $t \in \text{Disc}(x_n)^c$, $R(u_n)(s_n) = R(x_n)(t)$ by Lemma 8.4.1. Since $0 \leq s_n \leq 1$, $\{s_n\}$ has a convergent subsequence $\{s_{n_k}\}$. Let s' be the limit of that convergent subsequence. Since $r_{n_k}(s_{n_k}) = t$ for all n_k , we necessarily have $s' \in [s_-(t), s_+(t)]$. Since $\|R(u_n) - R(u)\| \rightarrow 0$, $R(x_{n_k})(t) = R(u_{n_k})(s_{n_k}) \rightarrow R(u)(s')$. Since we have already seen that $R(x_n)(t) \rightarrow R(x)(t)$, we must have $R(u)(s') = R(x)(t)$. Since $R(u)$ is constant on $[s_-(t), s_+(t)]$, we must have $R(u)(s) = R(x)(t)$ for all s with $s_-(t) \leq s \leq s_+(t)$.

(b) Since R maps D into D and C into C , $R(x)$ is right-continuous with left limits, while $R(u)$ is continuous. Given $t \in \text{Disc}(x)$, we can find $t_n \in \text{Disc}(x)^c$ with $t_n \uparrow t$. We can apply part (a) to obtain $R(u)(s_+(t_n)) = R(x)(t_n) \rightarrow R(x)(t-)$, but $s_+(t_n) \uparrow s_-(t)$, so that $R(u)(s_+(t_n)) \rightarrow R(u)(s_-(t))$. Hence, we have established the first claim: $R(u)(s_-(t)) = R(x)(t-)$. Similarly, we can find $t_n \in \text{Disc}(x)^c$ with $t_n \downarrow t$. Then we can apply part (a) again to obtain $R(u)(s_-(t_n)) = R(x)(t_n) \rightarrow R(x)(t)$. Since $s_-(t_n) \downarrow s_+(t)$, $R(u)(s_-(t_n)) \downarrow R(u)(s_+(t))$. Hence $R(x)(t) = R(u)(s_+(t))$ as claimed.

(c) We can apply Lemma 14.3.4 (a) in the book. Since the increment $x(t) - x(t-)$ is nonnegative in each component,

$$z(t) = z(t-) + x(t) - x(t-)$$

and $y(t) = y(t-)$. Similarly,

$$\phi(u)(s) = \phi(u)(s_-(t)) + u(s) - u(s_-(t))$$

and $\psi(u)(s) = \psi(u)(s_-(t))$ for $s_-(t) \leq s \leq s_+(t)$.

(d) We apply Lemma 14.3.4 (b) in the book. Each coordinate $\phi^i(u)$ and $\psi^i(u)$ is monotone in s over $[s_-(t), s_+(t)]$, so that the desired conclusion holds.

Proof of Theorem 8.4.2. (a) We combine parts (a)–(c) of Theorem 8.4.1 to get $(R(u), r)(s) \in \Gamma_{R(x)}$ for all s . Since R maps C into C , $(R(u), r)$ is continuous. Also r is nondecreasing with $r(0) = 0$ and $r(1) = T$ because $(u, r) \in \Pi_s(x)$. Finally, $(R(u), r)$ maps $[0, 1]$ onto $\Gamma_{R(x)}$ and $(R(u), v)$ is nondecreasing with respect to the order on $\Gamma_{R(x)}$ because the increments of $R(u)$ coincide with the increments of u over each discontinuity in x because $x \in D_+$, and (u, r) has these properties.

(b) We incorporate part (d) of Theorem 8.4.1 to get $R(u)$ monotone over $[s_-(t), s_+(t)] = r^{-1}(t)$ for each $t \in \text{Disc}(x) = \text{Disc}(R(x))$. This allows us to conclude that $(R(u), r) \in \Pi_w(R(x))$. ■

We now turn to the proof of Theorem 8.4.3. For the proof, we find it convenient to use a different class of approximating functions. Let D_l be the subset of all functions in D that (i) have only finitely many jumps and (ii) are continuous and piecewise linear in between jumps with only finitely many changes of slope. Let $D_{s,l} = D_s \cap D_l$.

Analogous to Theorem 6.2.2 in Section 6.2, we have the following result.

Lemma 8.4.4. (approximation of elements of D_s by elements of $D_{s,l}$) *For any $x \in D_s$, there exist $x_n \in D_{s,l}$ such that $\|x_n - x\| \rightarrow 0$ as $n \rightarrow \infty$.*

Proof. For $x \in D_s$ and $\epsilon > 0$ given, apply Theorem 6.2.2 in Section 6.2 to find $x_1 \in D_c$ (with only finitely many discontinuities) such that $\|x - x_1\| < \epsilon/4$. The function x_1 can have jumps of opposite sign, but the magnitude of the jumps in one of the two directions must be at most $\epsilon/2$. Form the desired function, say x_2 , from x_1 . Suppose that $\{t_1, \dots, t_k\} = \text{Disc}(x_1)$. Suppose that x_1 has one or more negative jump at t_j , none of which has

magnitude exceeding $\epsilon/2$. If x_1 has a negative jump at t_j in coordinate i for some i , then replace x_1^i over $[t_{j-1}, t_j]$ by the linear function connecting $x_1^i(t_{j-1})$ and $x_1^i(t_j)$. Similarly, if x_1 has one or more positive jumps at some t_j with all magnitudes less than $\epsilon/2$, then proceed as above. It is easy to see that $\text{Disc}(x_2) \subseteq \text{Disc}(x_1)$, $x_2 \in D_{s,l}$ and $\|x - x_2\| < \epsilon$. ■

We now show that limits of parametric representations are parametric representations when $\|x_n - x\| \rightarrow 0$.

Lemma 8.4.5. (limits of parametric representations) *If (i) $\|x_n - x\| \rightarrow 0$ as $n \rightarrow \infty$, (ii) $(u_n, r_n) \in \Pi_z(x_n)$ for each n , where $z = s$ or w , and (iii) $\|u_n - u\| \vee \|r_n - r\| \rightarrow 0$ as $n \rightarrow \infty$ where u and r are functions mapping $[0, 1]$ into $\mathbb{R}^k$ and $\mathbb{R}^1$, respectively, then $(u, r) \in \Pi_z(x)$ for the same z .*

Proof. Since (u, r) is the uniform limit of the continuous functions (u_n, r_n) , (u, r) is itself continuous. Since r is the limit of the nondecreasing functions r_n , r is itself nondecreasing. Since $r_n(0) = 0$ and $r_n(1) = T$ for all n , $r(0) = 0$ and $r(1) = T$. Since r is also nondecreasing and continuous, r maps $[0, 1]$ onto $[0, T]$. Pick any s with $0 < s < 1$. Then $r(s) = t$ for some t , $0 \leq t \leq T$, and $r_n(s) = t_n \rightarrow t$ as $n \rightarrow \infty$. Suppose that $(u_n, r_n) \in \Pi_s(x_n)$ for all n . That means that

$$u_n(s) = \alpha_n(s)x_n(t_n) + (1 - \alpha_n(s))x_n(t_n -)$$

for all n . Since $0 \leq \alpha_n(s) \leq 1$, there exists a convergent subsequence $\{\alpha_{n_k}(s)\}$ such that $\alpha_{n_k}(s) \rightarrow \alpha(s)$ as $n_k \rightarrow \infty$. At least one of the following three cases must prevail: (i) $t_{n_k} > t$ for infinitely many n_k , (ii) $t_{n_k} = t$ for infinitely many n_k and (iii) $t_{n_k} < t$ for infinitely many n_k . In case (i), we can choose a further subsequence $\{n_{k_j}\}$ so that $u_{n_{k_j}}(s) \rightarrow x(t)$; in case (ii), we can choose a further subsequence so that $u_{n_{k_j}}(s) \rightarrow \alpha(s)x(t) + [1 - \alpha(s)]x(t-)$; in case (iii) we can choose a further subsequence so that $u_{n_{k_j}}(s) \rightarrow x(t-)$. Since $u_n(s) \rightarrow u(s)$, the limit of the subsequence must be $u(s)$. Hence, $(u(s), r(s)) \in \Gamma_x$ for each s . Since (u, r) is continuous with $r(0) = 0$ and $r(1) = T$, (u, r) maps $[0, 1]$ onto Γ_x . Since (u_n, r_n) is monotone as a function from $[0, 1]$ to $(\Gamma_x, \leq)$ and $\|u_n - u\| \vee \|r_n - r\| \rightarrow 0$, (u, r) is monotone from $[0, 1]$ to $(\Gamma_x, \leq)$. Hence, $(u, r) \in \Pi_s(x)$. Finally, suppose that $(u_n, r_n) \in \Pi_w(x_n)$ for all n . By the result above applied to the individual coordinates, $(u^i(s), r(s)) \in \Gamma_{x^i}$ and thus $(u^i, r) \in \Pi_s(x^i)$ for each i , which implies that $(u, r) \in \Pi_w(x)$. ■

Proof of Theorem 8.4.3. For $x \in D_s$, apply Lemma 8.4.4 to find $x_n \in D_{s,l}$ such that $\|x_n - x\| \rightarrow 0$. Suppose that $(u, r) \in \Pi_w(x)$. Then it is possible to find u_n such that $(u_n, r) \in \Pi_w(x_n)$ and $\|u_n - u\| \rightarrow 0$. To do so, let $u_n(s_-(t)) = x_n(t-)$ and $u_n(s_+(t)) = x_n(t)$, where $[s_-(t), s_+(t)] = r^{-1}(t)$ for each $t \in \text{Disc}(x)$. If $t \in \text{Disc}(x_n)^c$, let $u_n(s) = u_n(s_+(t))$ for $s_-(t) \leq s \leq s_+(t)$; if $t \in \text{Disc}(x_n)$, define u_n so that $\|u_n - u\| \rightarrow 0$. Given that $(u_n, r) \in \Pi_w(x_n)$, we can apply mathematical induction over the finitely many time points such that x_n has a jump or a change of slope to show that $(R(u_n), r) \in \Pi_w(R(x_n))$ for each n . We use Lemma 14.3.4 of the book critically at this point to treat the discontinuity points of x_n in $D_{s,l}$. The continuous linear pieces between discontinuities can be treated by applying the rescaling property in Theorem 8.2.6 (iii) with $\beta = 1$ and $\eta = 0$. Finally, we apply Lemma 8.4.5 to deduce that $(R(u), r) \in \Pi_w(R(x))$. For that, we use the fact that $\|R(x_n) - R(x)\| \rightarrow 0$ and $\|R(u_n) - R(u)\| \rightarrow 0$.

8.5. M_1 Continuity Results

In this section we establish continuity and Lipschitz properties of the reflection map on $D \equiv D^k \equiv D([0, T], \mathbb{R}^k)$ with the M_1 topologies. Our first result establishes continuity of the reflection map R (for an arbitrary reflection matrix Q) as a map from (D, SM_1) to (D, L_1) , where L_1 is the topology on D induced by the L_1 norm

$$\|x\|_{L_1} \equiv \int_0^T \|x(t)\| dt. \quad (5.1)$$

Under a further restriction, the map from (D, WM_1) to (D, WM_1) will be continuous.

Recall that D_s is the subset of functions in D without simultaneous jumps of opposite sign in the coordinate functions; i.e., $x \in D_s$ if, for all $t \in (0, T)$, either $x(t) - x(t-) \leq 0$ or $x(t) - x(t-) \geq 0$, with the sign allowed to depend upon t . The subset D_s is a closed subset of D in the J_1 topology and thus a measurable subset of D with the SM_1 and WM_1 topologies (since the Borel σ -fields coincide). The proofs of the main theorems here appear in Section 6.2 of the Internet Supplement.

Theorem 8.5.1. (continuity with the SM_1 topology on the domain) *Suppose that $x_n \rightarrow x$ in (D, SM_1) .*

(a) *Then*

$$R(x_n)(t_n) \rightarrow R(x)(t) \quad \text{in } \mathbb{R}^{2k} \quad (5.2)$$

for each $t \in \text{Disc}(x)^c$ and sequence $\{t_n : n \geq 1\}$ with $t_n \rightarrow t$,

$$\sup_{n \geq 1} \|R(x_n)\| < \infty, \quad (5.3)$$

$$R(x_n) \rightarrow R(x) \quad \text{in} \quad (D, L_1) \quad (5.4)$$

and

$$\psi(x_n) \rightarrow \psi(x) \quad \text{in} \quad (D, WM_1). \quad (5.5)$$

(b) If in addition $x \in D_s$, then

$$\phi(x_n) \rightarrow \phi(x) \quad \text{in} \quad (D, WM_1), \quad (5.6)$$

so that

$$R(x_n) \rightarrow R(x) \quad \text{in} \quad (D, WM_1). \quad (5.7)$$

Proof. (a) We first prove (5.2). Since $x_n \rightarrow x$ in (D, SM_1) , we can find parametric representations $(u, r) \in \Pi_s(x)$ and $(u_n, r_n) \in \Pi_s(x_n)$ for $n \geq 1$ such that

$$\|u_n - u\| \vee \|r_n - r\| \rightarrow 0.$$

By Theorem 14.4.1 (a) in the book, $R(u)(s) = R(x)(t)$ for any $s \in [s_-(t), s_+(t)] \equiv r^{-1}(t)$, since $t \in \text{Disc}(x)^c$. Moreover, by Corollary 14.3.4 in the book, $t \in \text{Disc}(R(x))^c$. For any sequence $\{t_n : n \geq 1\}$ with $t_n \rightarrow t$, we can find another sequence $\{t'_n : n \geq 1\}$ such that $t'_n \rightarrow t$, $t'_n \in \text{Disc}(x_n)^c$ and $\|R(x_n)(t'_n) - R(x_n)(t_n)\| \rightarrow 0$ as $n \rightarrow \infty$. (Here we exploit the fact that $R(x_n) \in D$ for each n .) Consequently, $R(x_n)(t_n) \rightarrow R(x)(t)$ if and only if $R(x_n)(t'_n) \rightarrow R(x)(t)$. By Theorem 13.4.1 (a) again, $R(u_n)(s_n) = R(x)(t'_n)$ for any $s_n \in [s_-(t'_n), s_+(t'_n)] = r_n^{-1}(t'_n)$. Since $0 \leq s_n \leq 1$ for all n , any such sequence $\{s_n : n \geq 1\}$ has a convergent subsequence $\{s_{n_k} : k \geq 1\}$. Suppose that $s_{n_k} \rightarrow s'$ as $n_k \rightarrow \infty$. Since $t'_n \rightarrow t$ as $n \rightarrow \infty$ and $t'_{n_k} = r_{n_k}(s_{n_k}) \rightarrow r(s')$ as $n_k \rightarrow \infty$, we must have $s' \in [s_-(t), s_+(t)]$. Then, since $\|R(u_n) - R(u)\| \rightarrow 0$,

$$R(x_{n_k})(t'_{n_k}) = R(u_{n_k})(s_{n_k}) \rightarrow R(u)(s') = R(x)(t).$$

Since every subsequence of $\{R(x_n)(t'_n) : n \geq 1\}$ must have a convergent subsequence with the same limit, we must have $R(x_n)(t'_n) \rightarrow R(x)(t)$ as $n \rightarrow \infty$, which we have shown implies that $R(x_n)(t_n) \rightarrow R(x)(t)$ as $n \rightarrow \infty$, as claimed in (5.2). Next we establish (5.3). For any $x \in D$, $\|x\| \equiv \sup_{0 \leq t \leq T} \|x(t)\| < \infty$. Since $d_s(x_n, x) \rightarrow 0$, $\|x_n\| \rightarrow \|x\|$ as $n \rightarrow \infty$. Hence, it suffices to show that there is a constant K such that

$$\|R(x)\| \leq K\|x\| \quad \text{for all } x \in D,$$

but that follows from Theorem 13.2.5. We apply the bounded convergence theorem with (5.2) and (5.3) to establish (5.4). We now turn to (5.5). Since $\psi(x_n)$ and $\psi(x)$ are nondecreasing in each coordinate the pointwise convergence established in (5.2) actually implies WM_1 convergence in (5.5); see Corollary 12.5.1 in the book.

(b) First, we use the assumed convergence $x_n \rightarrow x$ in (D, SM_1) to pick $(u, r) \in \Pi_s(x)$ and $(u_n, r_n) \in \Pi_s(x_n)$, $n \geq 1$, with

$$\|u_n - u\| \vee \|r_n - r\| \rightarrow 0 .$$

Since R is continuous on (D, U) , we also have $\|R(u_n) - R(u)\| \rightarrow 0$. By part (a), we know that there is local uniform convergence of $R(x_n)$ to $R(x)$ at each continuity point of $R(x)$. Thus, by Theorem 12.5.1 (v) in the book, to establish $R(x_n) \rightarrow R(x)$ in (D, WM_1) , it suffices to show that

$$\lim_{\delta \downarrow 0} \overline{\lim}_{n \rightarrow \infty} w_s(R^i(x_n), t, \delta) = 0 \quad (5.8)$$

for each i , $1 \leq i \leq 2k$, and $t \in \text{Disc}(R(x))$, where

$$w_s(x, t, \delta) \equiv \sup\{\|x(t_2) - [x(t_1), x(t_3)]\| : (t_1, t_2, t_3) \in A(t, \delta)\} \quad (5.9)$$

for

$$A(t, \delta) \equiv \{(t_1, t_2, t_3) : (t - \delta) \vee 0 \leq t_1 < t_2 < t_3 \leq (t + \delta) \wedge T\} .$$

(Since we are considering the i^{th} coordinate function $R^i(x_n)$, the function x in (5.9) is real-valued here.) Suppose that (5.8) fails for some i and t . Then there exist $\epsilon > 0$ and subsequences $\{\delta_k\}$ and $\{n_k\}$ such that $\delta_k \downarrow 0$, $n_k \rightarrow \infty$ and

$$w_s(R^i(x_{n_k}), t, \delta_k) > \epsilon \quad \text{for all } \delta_k \text{ and } n_k .$$

That is, there exist time points t_{1,n_k} , t_{2,n_k} and t_{3,n_k} with

$$(t - \delta_k) \vee 0 \leq t_{1,n_k} < t_{2,n_k} < t_{3,n_k} \leq (t + \delta_k) \wedge T \quad (5.10)$$

and

$$\|R^i(x_{n_k})(t_{2,n_k}) - [R^i(x_{n_k})(t_{1,n_k}), R^i(x_{n_k})(t_{3,n_k})]\| > \epsilon . \quad (5.11)$$

Since the values $R^i(x_{n_k})(t)$ are contained in the values $R^i(u_{n_k})(s)$ where $(u_{n_k}, r_{n_k}) \in \Pi_s(x_{n_k})$, we can deduce that there are points s_{j,n_k} for $j = 1, 2, 3$ such that $0 \leq s_{1,n_k} < s_{2,n_k} < s_{3,n_k} \leq 1$, $r_{n_k}(s_{j,n_k}) = t_{j,n_k}$ for $j = 1, 2, 3$ and all n_k , and

$$\|R^i(u_{n_k})(s_{2,n_k}) - [R^i(u_{n_k})(s_{1,n_k}), R^i(u_{n_k})(s_{3,n_k})]\| > \epsilon . \quad (5.12)$$

By (5.10) and (5.12), there then exists a further subsequence $\{n'_k\}$ such that $t_{j,n'_k} \rightarrow t$ and $s_{j,n'_k} \rightarrow s_j$ as $n'_k \rightarrow \infty$ for $j = 1, 2, 3$, where $0 \leq s_1 \leq s_2 \leq s_3 \leq 1$, $r_{n'_k}(s_{j,n'_k}) \rightarrow r(s_j) = t$ and

$$\|R^i(u)(s_2) - [R^i(u)(s_1), R^i(u)(s_3)]\| \geq \epsilon > 0. \quad (5.13)$$

However, by Theorem 14.4.2 in the book, $(R(u), r) \in \Pi_w(R(x))$ since $x \in D_s$, so that $(R^i(u), r) \in \Pi_s(R^i(x))$. Hence $(R^i(u), r) \in \Pi_s(R^i(x))$. Since $R^i(u)$ is monotone on $[s_-(t), s_+(t)]$, (5.13) cannot occur. Hence (5.8) must in fact hold and $R^i(x_n) \rightarrow R^i(x)$ in (D, M_1) . Since that is true for all i , we must have $R(x_n) \rightarrow R(x)$ in (D, WM_1) . ■

Under the extra condition in part (b), the mode of convergence on the domain actually can be weakened. However, little positive can be said if only $x_n \rightarrow x$ in (D, WM_1) without $x \in D_s$; see Example 14.5.3 in the book.

Theorem 8.5.2. (continuity with the WM_1 topology on the domain) *If $x_n \rightarrow x$ in (D, WM_1) and $x \in D_s$, then (5.7) holds.*

The proof of Theorem 8.5.2 is more difficult. We now work towards its proof. By Theorem 8.4.3, R is Lipschitz on (D_s, WM_1) , but x_n need not be in D_s . We show that we can approximate x_n by elements of D_s .

We first restate Corollary 12.11.2 in the book as a lemma. It states that Convergence in WM_2 , which of course is implied by convergence in WM_1 , has the advantage that jumps in the converging functions must be inherited by the limit function.

Lemma 8.5.1. (inheritance of jumps) *If $x_n \rightarrow x$ in (D, WM_2) , $t_n \rightarrow t$ in $[0, T]$ and $x_n^i(t_n) - x_n^i(t_n-) \geq c > 0$ for all n , then $x^i(t) - x^i(t-) \geq c$.*

For $x \in D$ and $t \in \text{Disc}(x)$, let $\gamma(x, t)$ be the largest magnitude (absolute value) of the jumps in x at time t of opposite sign to the sign of the largest jump in x at time t . Let $\gamma(x)$ be the maximum of $\gamma(x, t)$ over all $t \in \text{Disc}(x)$. We apply Lemma 8.5.1 to establish the next result.

Lemma 8.5.2. *If $x_n \rightarrow x$ in (D, WM_1) , then*

$$\overline{\lim}_{n \rightarrow \infty} \gamma(x_n) \leq \gamma(x).$$

We only use the following consequence of Lemma 8.5.2.

Lemma 8.5.3. *If $x_n \rightarrow x$ in (D, WM_1) and $x \in D_s$, then $\gamma(x_n) \rightarrow 0$.*

We also use a generalization of Lemma 8.4.4 above, which is established in the same way.

Lemma 8.5.4. *For any $x \in D$, there exist $x_n \in D_{s,l}$ such that $\|x_n - x\| \rightarrow \gamma(x)$ as $n \rightarrow \infty$.*

We combine Lemmas 8.5.2 and 8.5.4 to obtain the tool we need.

Lemma 8.5.5. *If $x_n \rightarrow x$ in (D, WM_1) and $x \in D_s$, then there exists $x'_n \in D_{s,l}$ for $n \geq 1$ such that $\|x'_n - x_n\| \rightarrow 0$.*

Proof of Theorem 8.5.2. Given $x_n \rightarrow x$ in (D, WM_1) , apply Lemma 8.5.5 to find $x'_n \in D_{s,l}$ for $n \geq 1$ such that $\|x'_n - x_n\| \rightarrow 0$ as $n \rightarrow \infty$. Then, by the triangle inequality, Theorem 14.2.5 in the book and Lemma 8.5.3 above,

$$\begin{aligned} d_p(R(x_n), R(x)) &\leq d_p(R(x_n), R(x'_n)) + d_p(R(x'_n), R(x)) \\ &\leq \|R(x_n) - R(x'_n)\| + d_w(R(x'_n), R(x)) \\ &\leq K\|x_n - x'_n\| + Kd_w(x'_n, x). \end{aligned}$$

Since

$$\begin{aligned} d_p(x'_n, x) &\leq d_p(x'_n, x_n) + d_p(x_n, x) \\ &\leq \|x'_n - x_n\| + d_p(x_n, x) \\ &\rightarrow 0, \end{aligned}$$

$d_w(x'_n, x) \rightarrow 0$. Hence, $d_p(R(x_n), R(x)) \rightarrow 0$ as claimed. ■

Example 12.3.1 in the book shows that convergence $x_n \rightarrow x$ can hold in (D, WM_1) but not in (D, SM_1) even when $x \in D_s$. Thus Theorems 8.5.1 (a) and 8.5.2 cover distinct cases. An important special case of both occurs when $x \in D_1$, where D_1 is the subset of x in D with discontinuities in only one coordinate at a time; i.e., $x \in D_1$ if $t \in Disc(x^i)$ for at most one i when $t \in Disc(x)$, with the coordinate i allowed to depend upon t . In Section 6.7 it is shown that WM_1 convergence $x_n \rightarrow x$ is equivalent to SM_1 convergence when $x \in D_1$.

Just as with D_s above, D_1 is a closed subset of (D, J_1) and thus a Borel measurable subset of (D, SM_1) . Since $D_1 \subseteq D_s$, the following corollary to Theorem 8.5.2 is immediate.

Corollary 8.5.1. (common case for applications) *If $x_n \rightarrow x$ in (D, WM_1) and $x \in D_1$, then $R(x_n) \rightarrow R(x)$ in (D, WM_1) .*

We can obtain stronger Lipschitz properties on special subsets. Let D_+ be the subset of x in D with only nonnegative jumps, i.e., for which $x^i(t) - x^i(t-) \geq 0$ for all i and t . As with D_s and D_1 above, D_+ is a closed subset of (D, J_1) and thus a measurable subset of (D, SM_1) .

Theorem 8.5.3. (*Lipschitz properties*) *There is a constant K (the same as associated with the uniform norm from Theorem 8.2.5) such that*

$$d_s(R(x_1), R(x_2)) \leq K d_s(x_1, x_2) \quad (5.14)$$

for all $x_1, x_2 \in D_+$, and

$$d_p(R(x_1), R(x_2)) \leq d_w(R(x_1), R(x_2)) \leq K d_w(x_1, x_2) \leq K d_s(x_1, x_2) \quad (5.15)$$

for all $x_1, x_2 \in D_s$.

Proof. Given that $x \in D_+$, apply Theorem 14.4.2 (a) in the book to get $(R(u), r) \in \Pi_s(R(x))$ when $(u, r) \in \Pi_s(x)$. Then

$$\begin{aligned} d_s(R(x_1), R(x_2)) &= \inf_{\substack{(u'_i, r'_i) \in \Pi_s(R(x_i)) \\ i=1,2}} \{ \|u'_1 - u'_2\| \vee \|r_1 - r_2\| \} \\ &\leq \inf_{\substack{(u_i, r_i) \in \Pi_s(x_i) \\ i=1,2}} \{ \|\phi(u_1) - \phi(u_2)\| \vee \|r_1 - r_2\| \} \\ &\leq \inf_{\substack{(u_i, r_i) \in \Pi_s(x_i) \\ i=1,2}} \{ K \|u_1 - u_2\| \vee \|r_1 - r_2\| \} \\ &\leq K d_s(x_1, x_2) \end{aligned}$$

because $K \geq 1$. The other results are obtained in essentially the same way. Apply Theorem 14.4.3 in the book to get $(R(u), r) \in \Pi_w(R(x))$ when $(u, r) \in \Pi_w(x)$ and $x \in D_+$. When $x \in D_s$, apply Theorem 13.4.2 (b) to get $(R(u), r) \in \Pi_w(R(x))$ when $(u, r) \in \Pi_s(x)$. ■

We can actually do somewhat better than in Theorem 8.5.1 when the limit is in D_+ .

Theorem 8.5.4. (strong continuity when the limits is in D_+) *If*

$$x_n \rightarrow x \quad \text{in} \quad (D, SM_1), \quad (5.16)$$

where $x \in D_+$, then

$$R(x_n) \rightarrow R(x) \quad \text{in} \quad (D, SM_1). \quad (5.17)$$

Proof. Suppose that $x_n \rightarrow x$ in (D, SM_1) . By Theorem 8.5.1(a), we have $\psi(x_n) \rightarrow \psi(x)$ in (D, WM_1) . Since $x \in D_+$, $\psi(x) \in C$, by Corollary 14.3.5 in the book. Hence the WM_1 convergence is equivalent to uniform convergence; i.e.,

$$\psi(x_n) \rightarrow \psi(x) \quad \text{in } D([0, T], \mathbb{R}^k, U) .$$

We can then apply addition with equation (14.2.6) in the book to get

$$R(x_n) \rightarrow R(x) \quad \text{in } D([0, T], \mathbb{R}^{2k}, SM_1) . \quad \blacksquare$$

Our final result shows how the reflection map behaves as a function of the reflection matrix Q , as well as x , with the M_1 topologies.

Theorem 8.5.5. (continuity as a function of (x, Q)) *Suppose that $Q_n \rightarrow Q$ in $\mathcal{Q}$.*

(a) *If $x_n \rightarrow x$ in (D^k, WM_1) and $x \in D_s$, then*

$$R_{Q_n}(x_n) \rightarrow R_Q(x) \quad \text{in } (D^{2k}, WM_1) . \quad (5.18)$$

(b) *If $x_n \rightarrow x$ in (D^k, SM_1) and $x \in D_+$, then*

$$R_{Q_n}(x_n) \rightarrow R_Q(x) \quad \text{in } (D^{2k}, SM_1) . \quad (5.19)$$

Proof. We only prove the first of the two results, since the two proofs are essentially the same. If $x_n \rightarrow x$ in (D, WM_1) with $x \in D_s$, then we can find $x'_n \in D_{s,l}$ for $n \geq 1$ such that $\|x_n - x'_n\| \rightarrow 0$ by Lemma 8.5.5. By Theorem 14.2.5 in the book,

$$\|R_{Q_n}(x_n) - R_{Q_n}(x'_n)\| \leq K_n \|x_n - x'_n\| \rightarrow 0 \quad (5.20)$$

because $K_n \rightarrow K < \infty$. By Theorem 14.4.3 in the book, $(R_Q(u), r) \in \Pi_w(R(x))$ when $x \in D_s$. So, for any $\epsilon > 0$ given, let $(u, r) \in \Pi_w(x)$ and $(u_n, r_n) \in \Pi_w(x'_n)$ such that $\|u_n - u\| \vee \|r_n - r\| \leq \epsilon$. Then $(R_Q(u), r) \in \Pi_w(R_Q(x))$, $(R_{Q_n}(u_n), r_n) \in \Pi_w(R_{Q_n}(x'_n))$ for $n \geq 1$ and

$$\|R_{Q_n}(u_n) - R_Q(u)\| < K(\epsilon + \|Q_n - Q\|) \quad (5.21)$$

by Theorem 14.2.9 and equation (14.2.35) in the book, so that

$$R_{Q_n}(x'_n) \rightarrow R_Q(x) \quad \text{in } (D^{2k}, WM_1) . \quad (5.22)$$

Combining (5.20), (5.22) and the triangle inequality with the metric d_p , we obtain (5.18). $\blacksquare$

We can apply Section 6.9 to extend the continuity and Lipschitz results to the space $D([0, \infty), \mathbb{R}^k)$.

Theorem 8.5.6. (extension of continuity results to $D([0, \infty), \mathbb{R}^k)$) *The convergence-preservation results in Theorems 8.5.1, 8.5.2 and 8.5.4 and Corollary 8.5.1 extend to $D([0, \infty), \mathbb{R}^k)$.*

Proof. Suppose that $x_n \rightarrow x$ in $D([0, \infty), \mathbb{R}^k)$ with the appropriate topology and that $\{t_j : j \geq 1\}$ is a sequence of positive numbers with $t_j \in \text{Disc}(x)^c$ and $t_j \rightarrow \infty$ as $j \rightarrow \infty$. Then, $r_{t_j}(x_n) \rightarrow r_{t_j}(x)$ in $D([0, \infty), \mathbb{R}^k)$ with the same topology as $n \rightarrow \infty$ for each j , where r_t is the restriction map to $D([0, t], \mathbb{R}^k)$. Under the specified assumptions,

$$r_{t_j}(R(x_n)) = R_{t_j}(r_{t_j}(x_n)) \rightarrow R_{t_j}(r_{t_j}(x)) = r_{t_j}(R(x)) \quad (5.23)$$

in $D([0, t_j], \mathbb{R}^{2k})$ with the specified topology as $n \rightarrow \infty$ for each j , which implies that

$$R(x_n) \rightarrow R(x) \quad \text{in} \quad D([0, \infty), \mathbb{R}^{2k}) \quad (5.24)$$

with the same topology as in (5.23). ■

Theorem 8.5.7. (extension of Lipschitz properties to $D([0, \infty), \mathbb{R}^k)$) *Let $R : D([0, \infty), \mathbb{R}^k) \rightarrow D([0, \infty), \mathbb{R}^{2k})$ be the reflection map with function domain $[0, \infty)$ defined by Definition 8.2.1. Let metrics associated with domain $[0, \infty)$ be defined in terms of restrictions by (??) in Section 6.9. Then the conclusions of Theorems 8.2.5, 8.2.7 and 8.5.3 also hold for domain $[0, \infty)$.*

Proof. Apply Theorem 12.9.4 in the book. ■

8.6. Limits for Stochastic Fluid Networks

Nothing has been omitted from Section 14.6 of the book.

8.7. Queueing Networks with Service Interruptions

Nothing has been omitted from Section 14.7 of the book.

8.8. The Two-Sided Regulator

Nothing has been omitted from Section 14.8 of the book.

8.9. Existence of a Limiting Stationary Version

In this section, drawing on and extending Kella and Whitt (1996), we show that there exists a proper limiting stationary version of a reflected stochastic process under natural conditions. We establish existence and uniqueness of the limiting stochastic process, but we do not otherwise characterize the limiting marginal distribution on $\mathbb{R}^k$ or determine how to calculate it.

Our existence and uniqueness results with general initial conditions cover the case of the reflected Lévy process obtained as the heavy-traffic limit of the vector-valued buffer-content stochastic processes in a stochastic fluid network, as in Section 14.6 of the book, when the exogenous input processes at the different nodes are independent Lévy processes (i.e., processes with stationary independent increments) under a natural condition on the net input rates. We also obtain useful results about more general reflected processes without the independence conditions.

8.9.1. The Main Results

We are given a net-input stochastic process $\{X(t) : t \geq 0\}$ and the associated reflected content stochastic process

$$Z(t) \equiv \phi(X)(t) \equiv X(t) + (I - Q)Y(t), \quad t \geq 0, \quad (9.1)$$

where $Y \equiv \psi(X)$ is the minimal nondecreasing nonnegative stochastic process such that $Z \geq 0$, as in Definition 8.2.1. We want to consider the limiting behavior as $t \rightarrow \infty$. We want to determine conditions under which

$$(Z_s(t_1), \dots, Z_s(t_m)) \Rightarrow (Z_*(t_1), \dots, Z_*(t_m)) \quad \text{in } \mathbb{R}^{km} \quad \text{as } s \rightarrow \infty \quad (9.2)$$

for all positive integers m and any m time points t_i with $0 \leq t_1 < \dots < t_m$, where

$$Z_s(t) \equiv Z(s + t), \quad t \geq 0, \quad s \geq 0, \quad (9.3)$$

and the limiting stochastic process $Z_* \equiv \{Z_*(t) : t \geq 0\}$ is a *stationary stochastic process*, i.e., where

$$(Z_*(t_1 + h), \dots, Z_*(t_m + h)) \stackrel{d}{=} (Z_*(t_1), \dots, Z_*(t_m))$$

for all positive integers m , any m time points t_i with $0 \leq t_1 < \dots < t_m$ and all $h > 0$. We also want the limit process to be proper, i.e., we want to have

$$P(Z_*(t) < \infty) = 1 \quad \text{for all } t.$$

We then call the stochastic process Z_* the *limiting stationary version* of Z .

We first observe that convergence of the finite-dimensional distributions in (9.2) for processes Z_s defined as in (9.3) directly implies that the limit process Z_* is stationary.

Lemma 8.9.1. (stationarity from convergence) *If*

$$(Z(s + t_1), \dots, Z(s + t_m)) \Rightarrow (Z_*(t_1), \dots, Z_*(t_m)) \quad \text{in } \mathbb{R}^{km} \quad (9.4)$$

as $s \rightarrow \infty$ for all positive integers m and all m time points t_i with $0 \leq t_1 < \dots < t_m$, then Z_* is a stationary process.

Proof. If (9.4) holds, then

$$(Z(s + t_1 + h), \dots, Z(s + t_m + h)) \Rightarrow (Z_*(t_1 + u), \dots, Z_*(t_m + u)) \quad (9.5)$$

as $s \rightarrow \infty$ for any u , $0 \leq u \leq h$, because we can let $s' = s + h - u$, $t'_i = t_i + u$, $1 \leq i \leq m$, and let $s' \rightarrow \infty$ with (9.4). Hence the distribution of the random vector on the right in (9.5) must be independent of u . ■

In order to obtain a unique limiting stationary version, we will assume that the net-input process X has *stationary increments*, i.e., the joint distribution of the random vector

$$(X(t_1 + s) - X(u_1 + s), \dots, X(t_m + s) - X(u_m + s))$$

in $\mathbb{R}^{km}$ is independent of s for all positive integers m and all m -tuples of real numbers $(t_1, \dots, t_m)$ and $(u_1, \dots, u_m)$. We assume that X is defined on the whole real line $(-\infty, \infty)$. As a consequence,

$$X_s \equiv \{X(t + s) - X(s) : t \geq 0\} \quad (9.6)$$

has a distribution as a random element of D^k independent of s . We will also assume that X has *ergodic increment*, i.e., the increment $X(t + s) - X(s)$ have finite mean and

$$t^{-1}X(t) \rightarrow E[X(1) - X(0)] \quad \text{w.p.1} \quad \text{as } t \rightarrow \infty.$$

Here is our main result: In addition to the assumptions above, it depends on the special initial condition $X(0) = 0$, which forces $Z(0) = Y(0) = 0$. The proof of the following result and several others are given at the end of the section.

Theorem 8.9.1. (existence of a limiting stationary version) *If X has stationary ergodic increments with $X(0) = 0$ and*

$$((I - Q)^{-1}E[X(1) - X(0)])^i < 0, \quad 1 \leq i \leq k, \quad (9.7)$$

then (9.2) holds, i.e., the finite-dimensional distributions of Z_s in (9.3) converge as $s \rightarrow \infty$ to the finite-dimensional distributions of a proper stationary stochastic process Z_ .*

We now show the necessity of condition (9.7), leaving untouched the boundary case of equality. In particular, we show that a proper limit cannot exist if the strict inequality in (9.7) is reversed in any coordinate i . Indeed, then the i^{th} coordinate of the reflected process grows without bound.

Theorem 8.9.2. (necessity of the drift condition) *Suppose that*

$$t^{-1}X(t) \rightarrow x \quad \text{in } \mathbb{R}^k \quad \text{w.p.1 as } t \rightarrow \infty. \quad (9.8)$$

If

$$(I - Q)^{-1}x \leq 0, \quad (9.9)$$

then

$$t^{-1}Z(t) \rightarrow 0 \quad \text{as } t \rightarrow \infty \quad \text{w.p.1} \quad (9.10)$$

for Z in (9.1). On the other hand, if $((I - Q^{-1})x)^i > 0$ for some i , then

$$\liminf_{t \rightarrow \infty} t^{-1}Z^i(t) > 0 \quad \text{for that } i. \quad (9.11)$$

Proof. By Corollary 3.2.1 in the Internet Supplement, the SLLN in condition (9.8) implies the stronger FSLN

$$\mathbf{X}_n \rightarrow x\mathbf{e} \quad \text{in } D \quad \text{w.p.1}$$

for

$$\mathbf{X}_n(t) \equiv n^{-1}X(nt), \quad t \geq 0.$$

By Theorem 8.2.5,

$$\phi(\mathbf{X}_n) \rightarrow \phi(x\mathbf{e}) \quad \text{in } D \quad \text{w.p.1 as } n \rightarrow \infty.$$

However, condition (9.9) implies that $\phi(x\mathbf{e}) = \mathbf{0}$. Hence, (9.10) is obtained by applying the projection map $\pi_1(x) = x(1)$. Finally, we obtain (9.11) from (9.8) after noting from (9.1) that $(I - Q)^{-1}Z \geq (I - Q)^{-1}X$. ■

Theorem 8.9.1 does not cover all cases, because it requires the special initial condition $X(0) = 0$. However, we also obtain additional results with other initial conditions below. A difficulty occurs because in general the initial condition $X(0)$ and the remaining net-input process $\{X(t) - X(0) : t \geq 0\}$ are dependent. Hence, in general we cannot talk about the increments process as if it did not depend upon the initial condition. Nevertheless, we are able to obtain some positive results. We first establish a tightness result; see Section 11.6 of the book.

Theorem 8.9.3. (tightness under general initial conditions) *If X has stationary ergodic increments, and if condition (9.7) holds, then the family of random variables $\{Z(t) : t \geq 0\}$ is tight in $\mathbb{R}^k$.*

Since tightness in product spaces is equivalent to tightness of the components in each coordinate by Theorem 11.6.7 in the book, Theorem 8.9.3 implies the following.

Corollary 8.9.1. (tightness of the finite-dimensional distributions) *Under the conditions of Theorem 8.9.3, the family $\{Z_s(t_1), \dots, Z_s(t_m) : s \geq 0\}$ is tight in $\mathbb{R}^{km}$ for every positive integer m and m time points $0 \leq t_1 < \dots < t_m$.*

We can combine Prohorov's theorem (Theorem 11.6.1 in the book) with monotonicity to obtain the following result.

Corollary 8.9.2. (convergence of subsequences) *Under the conditions of Theorem 8.9.3, every subsequence $\{Z(t_k) : k \geq 1\}$ based on a sequence $\{t_k : k \geq 1\}$ of nonnegative numbers has a convergent subsequence $\{Z(t'_k) : k \geq 1\}$. If $Z(t_k) \Rightarrow L$ in $\mathbb{R}^k$ as $t_k \rightarrow \infty$, then*

$$Z_*(0) \leq_{st} L, \quad (9.12)$$

where Z_* is the stationary process obtained in Theorem 8.9.1 and

$$P(L^i < \infty) = 1, \quad 1 \leq i \leq k.$$

If we can conclude that the process Z gets arbitrarily close to the origin, then we can replace tightness in Theorem 8.9.3 with convergence.

Theorem 8.9.4. (convergence if the origin is approached) *If, in addition to the assumptions of Theorem 8.9.3, for any $\epsilon > 0$ there exists random time T_ϵ with*

$$P(T_\epsilon < \infty) = 1 \quad (9.13)$$

such that

$$\|Z(T_\epsilon)\| < \epsilon, \quad (9.14)$$

then the finite-dimensional distributions of Z_s in (9.3) converge as $s \rightarrow \infty$ to the finite-dimensional distributions of the limit process Z_* in Theorem 8.9.1.

We can obtain a stronger conclusion if the origin is actually hit for all initial positions.

Theorem 8.9.5. (coupling if the origin is always hit) *If, in addition to the assumptions of Theorem 8.9.3, for each initial value $X(0)$, there exists a random time T with $P(T < \infty) = 1$ such that $Z(T) = 0$, then the process $\{Z(t) : t \geq 0\}$ couples with the stationary version in finite time, so that*

$$\lim_{s \rightarrow \infty} Ef(Z_s) = Ef(Z_*)$$

for all measurable real-valued functions f on D^k .

However in general $\{Z(t) : t \geq 0\}$ need never visit a neighborhood of the origin.

Example 8.9.1. *The process Z need not visit a neighborhood of the origin.* To see that it is possible to have $Z(t) \neq (0, \dots, 0)$, and even $\|Z(t)\| > c > 0$ for some constant c , for all $t \geq 0$ under the conditions of Theorem 8.9.3, consider a two-dimensional case in which either $X^1(t + \epsilon) - X^1(t) > \delta\epsilon$ or $X^2(t + \epsilon) - X^2(t) > \delta\epsilon$ for all t , where ϵ and δ are small positive constants. For example, let

$$V^1(t) = \begin{cases} \delta, & 3k \leq t < 3k + 2 \\ -1, & 3k + 2 \leq t < 3k + 3 \end{cases}$$

and

$$V^2(t) = \begin{cases} \delta, & 3k + 1 \leq t < 3k + 3 \\ -1, & 3k \leq t < 3k + 1 \end{cases}$$

for all nonnegative integers k . Let U be uniformly distributed on $[0, 3]$. Then $\{V(t) : t \geq 0\} \equiv \{(V^1(t + U), V^2(t + U)) : t \geq 0\}$ is a stationary process on the positive half line, so that $X(t) \equiv \int_0^t V(u) du$ is a net input process with stationary increments. It is easy to see that the content process associated with $Q = 0$ never hits the origin after time 0, and yet for $\delta < 1/2$

it has a proper steady-state distribution. Indeed, eventually $Z(t)$ follows the deterministic trajectory with $Z(3k - U) = (2\delta, 0)$, $Z(3k + 1 - U) = (0, \delta)$ and $Z(3k + 2 - U) = (\delta, 2\delta)$. This steady-state trajectory is reached for

$$t \geq 3 \left(1 + \frac{\max\{Z^1(0), Z^2(0)\}}{1 - 2\delta} \right).$$

By an appropriate choice of units, the limiting trajectory falls outside any neighborhood of the origin.

We can also modify Example 8.9.1 to construct two stable content processes which differ only in their initial conditions but do *not* couple in finite time.

Example 8.9.2. *Failure to couple in finite time.* We modify Example 8.9.1 by letting $Q_{1,2}^t = P_{2,1}^t = \epsilon$ for $0 < \epsilon < \delta$. The content process now approaches the deterministic trajectory with $Z(3k - U) = (2\delta - \epsilon + \epsilon', 0)$, $Z(3k + 1 - U) = (0, \delta - \epsilon + \epsilon')$ and $Z(3k + 2 - U) = (\delta, 2\delta - \epsilon + \epsilon')$, where $\epsilon' = (2\delta^2 - \epsilon\delta)/(1 + \delta)$. However, unlike Example 8.9.1, the content process typically does not reach this cycle in finite time. Suppose one of the two content processes starts above another, where they have the same net input process X . They move together until they hit a boundary. However, when the lower process is on a boundary and the other is not, the other coordinate of the two processes moves away from each other at rate ϵ . Hence the processes cannot couple on any boundary, although they do get closer in an appropriate metric as they hit the boundaries.

Since many of the limiting net-input processes X will be Lévy processes (i.e., will have stationary independent increments), we now add the independent increments property.

Theorem 8.9.6. (existence and uniqueness for Lévy net-input processes with independent coordinate processes) *Suppose that $X \equiv (X^1, \dots, X^k)$ has mutually independent marginal processes X^i , $1 \leq i \leq k$, each with stationary and independent increments, $X(0)$ is proper and condition (9.7) holds. Then the limit (9.2) holds and the limit has the same distribution as the limit $Z_*(0)$ associated with $X(0) = 0$.*

As mentioned in the beginning of this section, Theorem 8.9.6 applies to the limit process in Section 8.6 when the scaled versions of the exogenous arrival process C converge to a Lévy process with mutually independent

coordinate processes, because the only stochastic component in the net-input process X^i is C^i . However, in general, Theorem 8.9.6 does not apply to the heavy-traffic limits for the queueing network in Section 8.7. It does in the special case in which the coordinate limit process $\mathbf{X}^i$ depends only on the limit of the scaled process associated with the i^{th} coordinate arrival process.

It remains to establish more general conditions under which the assumptions of Theorems 8.9.4 and 8.9.5 are satisfied. It also remains to find useful expressions for the limiting distributions. Explicit expressions for the Laplace transforms of non-product-form two-dimensional stationary buffer-content distributions of stochastic fluid networks with Lévy exogenous input processes have been determined by Kella and Whitt (1992a) and Kella (1993).

8.9.2. Proofs

We now provide the missing proofs for the results above. We first establish some bounds and inequalities to be used in the proofs. Let $D_{\downarrow}^k$ be the subset of nonnegative nonincreasing functions in D^k . As before, let $D_{\uparrow}^k$ be the subset of nonnegative nondecreasing functions in D^k .

Theorem 8.9.7. (bounds and inequalities for the reflection map) *Assume that $x_1, x_2 \in D$ with $x_2 - x_1 \in D_{\uparrow}^k$, $x_3 = x_1 + (I - Q)\psi(x_2)$ and $w \geq 0$ in $\mathbb{R}^k$. Then*

- (i) $\phi(x_2) \geq \phi(x_1)$,
- (ii) $\psi(x_1) - \psi(x_2) \in D_{\uparrow}^k$,
- (iii) $\psi(x_1) - \psi(x_2) \leq (I - Q)^{-1}(x_2 - x_1)$,
- (iv) $\psi(x_3) = \psi(x_1) - \psi(x_2)$,
- (v) $0 \leq (I - Q)^{-1}(\phi(x_2) - \phi(x_1)) \leq (I - Q)^{-1}(x_2 - x_1)$,
- (vi) $0 \leq 1(\phi(x_2) - \phi(x_1)) \leq 1(x_2 - x_1)$,
- (vii) $(I - Q)^{-1}(\phi(x_1 + w) - \phi(x_1)) \in D_{\downarrow}^k$,
- (viii) $1(\phi(x_1 + w) - \phi(x_1)) \in D_{\downarrow}^k$.

Proof. Parts (i) and (ii) follow for $x_1, x_2 \in D_c$ by induction from Corollary 14.3.2 and Lemma 14.3.3 in the book. They then follow for $x_1, x_2 \in D$ by taking limits: Given $x_1, x_2 \in D$ with $x_2 - x_1 \in D_\uparrow$, it is possible to find $x_{1,n}$ and $x_{2,n} \in D_c$ with $x_{2,n} - x_{1,n} \in D_\uparrow$ for all n and $\|x_{j,n} - x_j\| \rightarrow 0$ as $n \rightarrow \infty$ for $j = 1, 2$. Part (iii) follows from Theorem 14.2.4 in the book because

$$\eta_1(x_2 - x_1) = x_2 - x_1 \quad \text{for } x_2 - x_1 \in D_\uparrow.$$

Turning to (iv), note that

$$0 \leq \phi(x_3) = x_1 + (I - Q)(\psi(x_2) + \psi(x_3)) \quad (9.15)$$

and

$$0 \leq \phi(x_1) = x_3 + (I - Q)(\psi(x_1) - \psi(x_2)). \quad (9.16)$$

From (9.15) and minimality of $\psi(x_1)$, it follows that $\psi(x_1) \leq \psi(x_2) + \psi(x_3)$ for any choice of x_1 and x_2 . From (9.16) and minimality of $\psi(x_3)$, it follows that $\psi(x_3) \leq \psi(x_1) - \psi(x_2)$. Hence we must have $\psi(x_3) = \psi(x_1) + \psi(x_2)$ as claimed. Parts (v) –(viii) follow from the relations $(I - Q)^{-1}\phi(x) = (I - Q)^{-1}x + \psi(x)$ and $1(I - Q) \geq 0$, and Theorem 14.2.4 in the book. ■

We now apply Theorem 8.9.7 to determine the shape of several mean values as a function of time.

Corollary 8.9.3. (concavity of mean values) *If X has stationary increments with $X(0) = 0$, then the functions $((I - Q)^{-1}E\phi(X)(t))^i$, $E\psi^i(X)(t)$ and $1E\phi(X)(t)$ are concave functions of t for each i .*

Proof. Apply parts (vii), (ii) and (viii) of Theorem 8.9.7, respectively. We will only prove the first result because the three proofs are essentially the same. It suffices to show that

$$((I - Q)^{-1}E[\phi(X)(t + s) - \phi(X)(s)])^i$$

is nonincreasing in s for all t , but that follows from Theorem 8.9.7(vii), because $\phi(X)(t + s)$ is distributed as the reflection of $X_s(t) \equiv X(s + t) - X(s)$ starting at $\phi(X)(s)$ evaluated at t , while $\phi(X)(s)$ is distributed as the reflection of X_s starting at 0 evaluated at t , since the law of X_s is independent of s . ■

We say that a real-valued function f on $\mathbb{R}_+$ is *subadditive* if

$$f(t_1 + t_2) \leq f(t_1) + f(t_2)$$

for all $t_1, t_2 \in \mathbb{R}_+$. We say that an $\mathbb{R}^k$ -valued stochastic process $\{X(t) : t \geq 0\}$ is *stochastically increasing and subadditive* (SIS) if

$$Ef(X(t_1 + t_2)) \leq Ef(X(t_1)) + Ef(X(t_2))$$

for all nondecreasing subadditive real-valued functions f on $\mathbb{R}^k$.

Corollary 8.9.4. (SIS property) *If X has stationary increments with $X(0) = 0$, then $(I - Q)^{-1}Z$ and $1Z$ are stochastically increasing and subadditive stochastic processes.*

Proof. Since the two results are proved similarly, we only prove the first. Let

$$\tilde{Z}_{s_1, s_2}(t) \equiv (I - Q)^{-1}Z(t)$$

with Z having initial value $Z(s_1)$ and net input $X_{s_2}(t) \equiv X(s_2 + t) - X(s_2)$, $t \geq 0$, where $0 \leq s_1 \leq s_2$. By Theorem 8.9.7(vii),

$$\tilde{Z}_{s, s}(t) - \tilde{Z}_{0, s}(t) \leq \tilde{Z}_{s, s}(0) - \tilde{Z}_{0, s}(0) = \tilde{Z}_{s, s}(0)$$

for all $s, t \geq 0$, or

$$\tilde{Z}_{0, 0}(s + t) = \tilde{Z}_{s, s}(t) \leq \tilde{Z}_{0, s}(t) + \tilde{Z}_{s, s}(0) ,$$

so that, for any subadditive function f ,

$$\begin{aligned} E[f(\tilde{Z}_{0, 0}(s + t))] &\leq E[f(\tilde{Z}_{0, s}(t) + \tilde{Z}_{s, s}(0))] \\ &\leq E[f(\tilde{Z}_{0, s}(t))] + E[f(\tilde{Z}_{s, s}(0))] \\ &\leq E[f(Z_{0, 0}(t))] + E[f(Z_{0, 0}(s))] , \end{aligned}$$

with the last line holding because there is equality in distribution for the respective terms. ■

A key to establishing the important Theorems 14.8.1 and 14.8.6 in the book is the following stochastic increasing property, which we deduce from Theorem 8.9.7.

Theorem 8.9.8. (stochastic increasing starting empty) *If X has stationary increments and $X(0) = 0$, then the family of processes $\{Z_s : s \geq 0\}$ in (9.3) is stochastically increasing in s , i.e.,*

$$Ef(Z_{s_1}) \leq Ef(Z_{s_2})$$

for $0 \leq s_1 < s_2$ and all bounded measurable nondecreasing real-valued functions f on $D \equiv D([0, \infty), \mathbb{R}^k)$, using the componentwise order on D .

Proof. Let $\hat{Z}_s(t)$ ($Z_s(t)$) be the content with $Z(0) = 0$ ($Z(0) = Z(s)$) and input increments from X_s in equation (14.8.6) in the book. Then, for $0 \leq s_1 < s_2$,

$$\hat{Z}_{s_2-s_1}(s_1+t) \leq Z_{s_2-s_1}(s_1+t) \quad \text{for all } t \geq 0 \quad \text{w.p.1,}$$

by Theorem 8.9.7 because $\hat{Z}_{s_2-s_1}(0) \equiv 0 \leq Z_{s_2-s_1}(0) \equiv Z(s_2-s_1)$ and both processes have the common input increments from X_s . Hence,

$$Ef(\hat{Z}_{s_2-s_1}) \leq Ef(Z_{s_2-s_1})$$

for all nondecreasing bounded measurable real-valued functions f on D , using the usual componentwise order. However, since $X_s \stackrel{d}{=} X$,

$$\{\hat{Z}_{s_1-s_1}(s_1+t) : t \geq 0\} \stackrel{d}{=} \{Z(s_1+t) : t \geq 0\} \equiv Z_{s_1}$$

and

$$\{Z_{s_2-s_1}(s_1+t) : t \geq 0\} = \{Z(s_2+t) : t \geq 0\} \equiv Z_{s_2}.$$

These last three relations combine to establish the desired conclusion. ■

We use the following result to establish Theorem 14.8.3 in the book.

Theorem 8.9.9. (tightness solidarity) *Suppose that X has stationary increments. Then $\{Z(t) : t \geq 0\}$ is tight for all proper distributions of $X(0)$ if and only if it is tight for any one.*

Proof. Note that $\{Z(t) : t \geq 0\}$ is tight if and only if $\{(I-Q)^{-1}Z(t) : t \geq 0\}$ is tight. By Theorem 8.9.7, the processes $(I-Q)^{-1}Z(t)$ starting at $X(0)$ and 0, with common increments from X , differ by at most $(I-Q)^{-1}\|X(0)\|$. Hence they are tight or non-tight together. Hence, the tightness of the process with one proper initial condition implies the tightness of the process starting at 0. Then the tightness of the process starting at 0 implies the tightness of any other process with another initial condition. ■

The key to our tightness results, and thus also our convergence results, is our ability to bound the marginal processes Z^i associated with a k -dimensional reflected process $Z \equiv (Z^1, \dots, Z^k)$ by related well-studied and well-understood one-dimensional reflections. For that purpose, we have the following bounds.

Theorem 8.9.10. (one-dimensional reflection bounds) *For any $x \in D^k$ and $Q \in \mathcal{Q}$,*

$$\psi_1((I-Q)^{-1}x) \leq \psi(x) \leq (I-Q)^{-1}\psi_1(x) \quad (9.17)$$

and

$$\phi_1((I - Q)^{-1}x) \leq (I - Q)^{-1}\phi(x) \leq (I - Q)^{-1}\phi_1(x) , \quad (9.18)$$

where $(\psi_1, \phi_1) : D^k \rightarrow D^{2k}$ with

$$(\psi_1(x)^i, \phi_1(x)^i) \equiv (\hat{\psi}_1(x^i), \hat{\phi}_1(x^i)), \quad 1 \leq i \leq k ,$$

and $(\hat{\psi}_1, \hat{\phi}_1) : D \rightarrow D^2$ being the one-dimensional reflection map, i.e.,

$$\hat{\phi}_1(x^i) \equiv x^i + \hat{\psi}_1(x^i) \quad (9.19)$$

and

$$\hat{\psi}_1(x^i) \equiv - \inf_{0 \leq s \leq t} \{x_i(s)^-\}, \quad t \geq 0 . \quad (9.20)$$

Proof. For the upper bounds, note that

$$\phi_1(x) = x + \psi_1(x) = x + (I - Q)(I - Q)^{-1}\psi_1(x) .$$

By the minimality of $\psi(x)$ in the definition of (ψ, ϕ) ,

$$\psi(x) \leq (I - Q)^{-1}\psi_1(x) .$$

Therefore,

$$(I - Q)^{-1}\phi(x) = (I - Q)^{-1}x + \psi(x) \leq (I - Q)^{-1}x + (I - Q)^{-1}\psi_1(x) = (I - Q)^{-1}\phi_1(x) .$$

Similarly, for the lower bound,

$$\phi_1((I - Q)^{-1}x) = (I - Q)^{-1}x + \psi_1((I - Q)^{-1}x) \quad (9.21)$$

and

$$(I - Q)^{-1}\phi(x) = (I - Q)^{-1}x + \psi(x) .$$

Since $(I - Q)^{-1}\phi(x) \geq 0$, we can apply the minimality of ψ_1 in (9.21) to deduce that

$$\psi_1((I - Q)^{-1}x) \leq \psi(x)$$

and

$$\phi_1((I - Q)^{-1}x) \leq (I - Q)^{-1}\phi(x) . \quad \blacksquare$$

In order to apply the one-dimensional reflection bounds in Theorem 8.9.10, we need to have a net input process X with negative drift in each coordinate. However, from (9.7), we only have X such that $(I - Q)^{-1}X$ has negative drift in each coordinate. We now show that, given X such that $(I - Q)^{-1}X$ has negative drift, we can bound $(I - Q)^{-1}\phi(X)$ above by $(I - Q)^{-1}\phi(X_y)$, where $X_y(t) \equiv X(t) - yt$, $t \geq 0$ and X_y has negative drift in each coordinate.

Theorem 8.9.11. (upper bound with negative drift) *Let X be a random element of D^k with stationary increments such that*

$$E[X(1) - X(0)] = x \quad \text{and} \quad ((I - Q)^{-1}x)^i < 0, \quad 1 \leq i \leq k.$$

For any $y \in \mathbb{R}^k$ with $y^i > x^i$ and $((I - Q)^{-1}y)^i < 0$, $1 \leq i \leq k$ (there necessarily is one), let

$$X_y(t) \equiv X(t) - yt, \quad t \geq 0.$$

Then X_y has stationary increments (and ergodic increments if X does) with

$$E[X_y(1) - X_y(0)]^i = x^i - y^i < 0, \quad 1 \leq i \leq k,$$

and

$$(I - Q)^{-1}\phi(X) \leq (I - Q)^{-1}\phi(X_y). \quad (9.22)$$

Proof. Only the final conclusion (9.22) requires discussion. Let e be the identity map, i.e., $e(t) = t$, $t \geq 0$. Recall that

$$\begin{aligned} \phi(X)(t) &\equiv X(t) + (I - Q)\psi(X)(t) \\ \phi(X_y)(t) &\equiv X(t) - yt + (I - Q)\psi(X_y)(t) \\ \phi(ye)(t) &\equiv yt + (I - Q)\psi(ye)(t), \quad t \geq 0. \end{aligned}$$

First, since $(I - Q)^{-1}y \leq 0$, it is easy to see that

$$\phi(ye)(t) = 0 \quad \text{and} \quad \psi(ye)(t) = -(I - Q)^{-1}yt.$$

Then

$$\phi(X_y)(t) \equiv \phi(X_y)(t) + \phi(ye)(t) = X(t) + (I - Q)(\psi(X_y)(t) + \psi(ye)(t)).$$

By the minimality of $\psi(X)$,

$$\psi(X) \leq \psi(X_y) + \psi(ye)$$

and

$$\begin{aligned} (I - Q)^{-1}\phi(X) &= (I - Q)^{-1}X + \psi(X) \\ &\leq (I - Q)^{-1}X + \psi(X_y) + \psi(ye) = (I - Q)^{-1}\phi(X_y). \quad \blacksquare \end{aligned}$$

We now state the classical one-dimensional result, which depends on the fact that the reflected content $\phi(X)(t)$ has the same distribution as the supremum of the time-reversed net-input process for each t (but not for multiple t).

Theorem 8.9.12. (classical one-dimensional result) *If X is a real-valued stochastic process with stationary increments such that*

$$X_r(t) \equiv -X(-t) \rightarrow -\infty$$

as $t \rightarrow \infty$ and $X(0)$ is proper, then there exists a proper random variable L such that

$$\phi(X)(t) \Rightarrow L \quad \text{in } \mathbb{R} \quad \text{as } t \rightarrow \infty.$$

Proof. First assume that $X(0) = 0$. Given the time reversed process $X_r(t) \equiv -X(-t)$, $t \geq 0$, note that

$$\phi(X)(t) \stackrel{d}{=} X_r^\uparrow(t) \quad \text{for each } t \geq 0.$$

Since $X_r(t) \rightarrow -\infty$ as $t \rightarrow \infty$ and $X_r \in D$,

$$X_r^\uparrow(t) \rightarrow X_r^\uparrow(\infty) < \infty \quad \text{as } t \rightarrow \infty \quad \text{w.p.1.}$$

Hence the desired conclusion holds with the proper limit $L \stackrel{d}{=} X_r^\uparrow(\infty)$. Now suppose that $X(0) \neq 0$. Since $X(t) \rightarrow -\infty$ w.p.1, the processes $Z(t)$ starting at 0 and $X(0)$, with common net input process X , couple w.p.1. Hence we can invoke Theorem 8.9.5. ■

We now provide the missing proofs of theorems earlier in this section.

Proof of Theorem 8.9.1. By Theorem 8.9.8, the family of processes Z_s in (9.3) are stochastically increasing in s . Consequently, the finite-dimensional distributions of Z_s are stochastically increasing in s . The cumulative distribution functions (cdf's) of $(Z_s(t_1), \dots, Z_s(t))$ in $\mathbb{R}^{km}$ thus converge as $s \rightarrow \infty$ to a possibly improper cdf; e.g., see Chapter VIII of Feller (1971). It thus suffices to show that $\{Z^i(t) : t \geq 0\}$ is tight for each i , for which it suffices to show that $\{((I - Q)^{-1}Z(t))^i : t \geq 0\}$ is tight for each i . (The tightness implies that the limiting cdf is proper.) By Theorem 8.9.11, we can bound $(I - Q)^{-1}\phi(X)$ above by $(I - Q)^{-1}\phi(X_y)$, where $X_y(t) \equiv X(t) - yt$ for appropriate $y \in \mathbb{R}^k$ and

$$-\infty < E[X_y^i(1) - X_y^i(0)] < 0 \quad \text{for all } i. \quad (9.23)$$

By (9.18) in Theorem 8.9.10, we can bound $(I - Q)^{-1}\phi(X_y)$ above by $(I - Q)^{-1}\phi_1(X_y)$, where ϕ_1 is the vector of one-dimensional reflection maps. Hence it suffices to show that $\{\hat{\phi}_1(X_y^i(t)) : t \geq 0\}$ is tight for each i , where

$\hat{\phi}_1$ is the one-dimensional reflection map in (9.19). However, $\hat{\phi}_1(X_y^i)(t)$ converges to a proper limit by Theorem 8.9.12. The condition $-X(-t) \rightarrow -\infty$ in Theorem 8.9.12 holds for X_y by virtue of (9.23) and that fact that $\{-X(t)\}$ is a process with stationary ergodic increments (Stationarity and metric transitivity are invariant under time reversal, and ergodicity is equivalent to metric transitivity.) The assumptions imply that

$$-t^{-1}X_y^i(-t) \rightarrow E[X_y^i(1) - X_y^i(0)] \quad \text{as } t \rightarrow \infty \quad \text{w.p.1}$$

for each i , which implies that $-X_y^i(-t) \rightarrow -\infty$ w.p.1 as $t \rightarrow \infty$ for each i . ■

Proof of Theorem 8.9.3. By Theorem 8.9.1, we have convergence to a proper limit L for the process $\{Z_0(t) : t \geq 0\}$ starting from the origin. By the continuous mapping theorem,

$$(I - Q)^{-1}Z_0(t) \Rightarrow (I - Q)^{-1}L \quad \text{as } t \rightarrow \infty.$$

If $X(0)$ is proper, then so is $X(0)^+ \equiv (X^1(0)^+, \dots, X^k(0)^+)$. Then, from Theorem 8.9.7(i) and (v),

$$0 \leq (I - Q)^{-1}Z_{X(0)}(t) \leq (I - Q)^{-1}Z_{X(0)^+}(t) \leq (I - Q)^{-1}Z_0(t) + (I - Q)^{-1}X(0)^+,$$

where here $Z_w(t)$ denotes the process governed by X with initial position w . Hence

$$\begin{aligned} P(|((I - Q)^{-1}Z_{X(0)}(t))^i| > 2K) &\leq P(|(I - Q)^{-1}Z_0(t))^i| > K) \\ &\quad + P(|(I - Q)^{-1}X(0)^+)^i| > K), \end{aligned}$$

so that the tightness holds by the results above. ■

Proof of Corollary 8.9.2. We can combine Prohorov's theorem (Theorem 11.6.1 in the book) with monotonicity. By Theorem 8.9.7,

$$Z_0(t) \leq Z_{X(0)}(t) \quad \text{for all } t. \quad (9.24)$$

Since $Z_0(t) \Rightarrow Z_*(0)$ by Theorem 8.9.1, (9.12) must hold. (Stochastic order on $\mathbb{R}^k$ is preserved under weak convergence.) ■

Proof of Theorem 8.9.4. Since (9.24) holds and

$$(I - Q)^{-1}(Z_{X_0} - Z_0) \in D_{\downarrow}^k,$$

by Theorem 8.9.7(vii),

$$0 \leq (I - Q)^{-1}(Z_{X(0)}(t) - Z_0(t)) \leq (I - Q)^{-1}1\epsilon \quad \text{for all } t \geq T_{\epsilon}.$$

Since $Z_0(t) \Rightarrow L$ as $t \rightarrow \infty$ by Theorem 8.9.1 and ϵ is arbitrary, we must have $Z_{X(0)}(t) \Rightarrow L$ too. ■

Proof of Theorem 8.9.5. The processes starting at 0, $X(0)$ or $Z_*(0)$ can all be given a common net input process $X(t) - X(0)$, $t \geq 0$. Hence, they all must couple when the process starting at $Z(0) \vee Z^*(0)$ first hits the origin. ■

In preparation for the proof of Theorem 8.9.6, we now establish a property of the limiting distribution in the one-dimensional case when X is a Lévy process.

Theorem 8.9.13. (mass near the origin) *If, in addition to the assumptions of Theorem 8.9.12, the one-dimensional net-input process X has independent increments, then*

$$P(L < \epsilon) > 0 \quad \text{for all } \epsilon > 0,$$

where L is the limiting random variable.

Proof. Consider the time reversed process X_r defined in Theorem 8.9.12. It suffices to show that $P(X_r^\uparrow(\infty) < \epsilon) > 0$. Suppose not. Then $P(X_r^\uparrow(\infty) \geq \epsilon) = 1$, which implies that $P(T_\epsilon < \infty) = 1$, where

$$T_\epsilon = \inf\{t > 0 : X_r(t) \geq \epsilon\}.$$

Using the regeneration property associated with the stationary independent increments, that in turn implies that

$$\limsup_{t \rightarrow \infty} X_r(t) = +\infty \quad \text{w.p.1},$$

which contradicts the limit $X_r(t) \rightarrow -\infty$ w.p.1. Hence we must have $P(X_r^\uparrow(\infty) < \epsilon) > 0$ for all $\epsilon > 0$ as claimed. ■

Proof of Theorem 8.9.6. The conditions allow us to apply Theorem 8.9.4. Theorems 8.9.10 and 8.9.11 allow us to bound the process $(I - Q)^{-1}\phi(X)(t)$ above by $(I - Q)^{-1}\phi_1(X_y)(t)$, as in the proof of Theorem 8.9.1. However, $\phi_1(X_y)$ has mutually independent coordinate processes. Let L^i be the limit random variable for the one-dimensional process associated with $\phi_1(X_y)$ and coordinate i . Since, for any $\epsilon > 0$,

$$P(L^1 \leq \epsilon, \dots, L^k \leq \epsilon) = \prod_{i=1}^k P(L^i \leq \epsilon) > 0$$

by the independence and Theorem 8.9.13 we must have $P(T_\epsilon < \infty) = 1$ for the random time T_ϵ in Theorem 8.9.4. ■

As mentioned earlier, Theorem 8.9.6 applies to the limit process in Section 14.6 in the book when the scaled versions of the exogenous arrival process C converge to a Lévy process with mutually independent coordinate processes, because the only stochastic component in the net-input process X^i is C^i . However, in general, Theorem 8.9.6 does not apply to the heavy-traffic limits for the queueing network in Section 14.7 of the book. It does in the special case in which the coordinate limit process $\mathbf{X}^i$ depends only on the limit of the scaled process associated with the i^{th} coordinate arrival process.

Chapter 9

Nonlinear Centering and Derivatives

9.1. Introduction

In this chapter we continue to study the useful functions introduced in Section 3.5 of the book and investigated in Chapter 13 of the book. Now we consider supremum, reflection and inverse maps with nonlinear centering.

Following Mandelbaum and Massey (1995), we identify the limit of the properly scaled function as a derivative. We also show how the convergence-preservation results for the reflection map can be applied to establish heavy-traffic limits for nonstationary queues.

To explain the derivative representation, recall that our previous results on the preservation of convergence with linear centering started with the assumed convergence

$$c_n(x_n - e) \rightarrow y \quad \text{in } D, \quad (1.1)$$

where $c_n \rightarrow \infty$ and e is the identity function, i.e., $e(t) = t$, $t \geq 0$. Given (1.1), we found conditions under which

$$c_n(\phi(x_n) - e) \rightarrow z \quad \text{in } D \quad (1.2)$$

for various functions ϕ and we identified the limit z . We also obtained some extensions in which the linear centering function e in (1.1) is replaced by a nonlinear function x ; i.e., instead of (1.1), we assumed that

$$c_n(x_n - x) \rightarrow y \quad \text{in } D \quad \text{as } n \rightarrow \infty, \quad (1.3)$$

where $c_n \rightarrow \infty$. In particular, see Theorems 13.3.2, 13.7.2 and 13.7.4 and Corollaries 13.4.1, 13.7.1 and 13.7.2 in the book. We now want to obtain some further results of this kind.

Given (1.3), we have as a consequence

$$x_n \rightarrow x \quad \text{in } D. \quad (1.4)$$

Hence, for any continuous function ϕ , we have

$$\phi(x_n) \rightarrow \phi(x) \quad \text{in } D. \quad (1.5)$$

Thus we want to find functions $z \in D$ and regularity conditions such that

$$c_n(\phi(x_n) - \phi(x)) \rightarrow z \quad \text{in } D. \quad (1.6)$$

The previous results with centering by e were of this form, where $\phi(x) = x = e$. The M topologies play an important role, because the limit z in (1.6) may have discontinuities even when y , x and x_n are all continuous functions.

In a probability context, (1.6) is interesting because it corresponds to a FCLT refinement to a nonlinear FLLN. We may have scaled stochastic processes $\{X_n(t) : t \geq 0\}$ which obey a nonlinear FWLLN of the form

$$X_n \Rightarrow x \quad \text{in } D, \quad (1.7)$$

where x is a nonlinear deterministic function, and a FCLT refinement of the form

$$c_n(X_n - x) \Rightarrow Y \quad \text{in } D, \quad (1.8)$$

where $c_n \rightarrow \infty$. From the FWLLN (1.7) it follows directly that

$$\phi(X_n) \Rightarrow \phi(x) \quad \text{in } D \quad (1.9)$$

for a continuous function ϕ . Our goal is to establish the FCLT refinement of (1.9), i.e.,

$$c_n(\phi(X_n) - \phi(x)) \Rightarrow Z \quad \text{in } D. \quad (1.10)$$

As before, (1.10) follows from (1.8) when (1.6) follows from (1.3). Hence we focus on obtaining (1.6) from (1.3).

It is interesting that, under regularity conditions, z in (1.6) can be thought of as a derivative of the map ϕ , in particular, a directional derivative of ϕ in the direction y , evaluated at x . To see that, it is convenient to index the functions by ϵ in such a way that x_n becomes x_ϵ and c_n becomes ϵ^{-1} . (That is without loss of generality.) Then (1.3) is equivalent to

$$\epsilon^{-1}(x_\epsilon - x) \rightarrow y \quad \text{as } \epsilon \downarrow 0. \quad (1.11)$$

Without being too precise, we can rewrite (1.11) as

$$x_\epsilon = x + \epsilon y + o(\epsilon) \quad \text{as } \epsilon \downarrow 0. \quad (1.12)$$

Now, assuming that the function $\phi : D \rightarrow D$ satisfies

$$\phi(\tilde{x}_\epsilon + o(\epsilon)) - \phi(\tilde{x}_\epsilon) = o(\epsilon) \quad \text{as } \epsilon \downarrow 0 \quad (1.13)$$

for any $\tilde{x}_\epsilon$ with $\tilde{x}_\epsilon \rightarrow x$ in D as $\epsilon \downarrow 0$ (which is not automatic), we have

$$\phi(x_\epsilon) = \phi(x + \epsilon y) + o(\epsilon) \quad \text{as } \epsilon \downarrow 0 \quad (1.14)$$

and, given the ϵ -analog of (1.6),

$$\phi(x + \epsilon y) = \phi(x) + \epsilon z + o(\epsilon) \quad \text{as } \epsilon \downarrow 0. \quad (1.15)$$

From (1.15), it is evident that z can be given the directional derivative interpretation. Moreover, (1.14) and (1.15) together imply that

$$\epsilon^{-1}(\phi(x_\epsilon) - \phi(x)) \rightarrow z \quad \text{as } \epsilon \downarrow 0. \quad (1.16)$$

Equivalently, (1.3), (1.13) and (1.16) imply the desired (1.6).

Here is how the present chapter is organized: In Section 2 we investigate when the convergence-preservation question (when (1.3) implies (1.6)) can be reduced to the derivative determination in (1.15). Unfortunately, we are not able to show that this can be done as generally as we would like. This step seems to be the weak link in our analysis in this chapter. Hopefully future research will provide further insights.

In Sections 9.3 – 9.5 we determine sufficient conditions for the derivatives of the supremum and reflection maps to exist and determine their form. As should be anticipated from Chapter 13 in the book, the reflection derivative can be expressed in terms of the supremum derivative. The M_1 topology plays an important role even if x and y in (1.3) are both continuous.

In Section 9.6 we apply the derivative calculation and convergence-preservation results for the reflection map to establish heavy-traffic limits for nonstationary queues. For example, these results cover the $M_t/M_t/1$ queue with time-dependent arrival and service rates.

Finally, in Section 9.7 we consider the derivative of the inverse map.

9.2. Nonlinear Centering and Derivatives

In this section we investigate when the desired convergence-preservation (when (1.11) implies (1.16)) can be deduced by determining the derivative

via (1.15). For any function $\phi : D \rightarrow D$, a general approach to establish the desired limit (1.16) for $\phi(x_\epsilon)$ is to exploit the triangle inequality:

$$\begin{aligned} d(\epsilon^{-1}[\phi(x_\epsilon) - \phi(x)], z) &\leq d(\epsilon^{-1}[\phi(x + \epsilon y) - \phi(x)], z) \\ &\quad + d(\epsilon^{-1}[\phi(x_\epsilon) - \phi(x)], \epsilon^{-1}[\phi(x + \epsilon y) - \phi(x)]) \end{aligned}$$

for an appropriate metric d . A limit for the first term in (2.1) as $\epsilon \downarrow 0$ identifies z as the derivative of ϕ in the direction y evaluated at x . In addition to establishing the existence of this derivative, we must also show that the second term in (2.1) converges to 0 as $\epsilon \downarrow 0$. Surprisingly, the second term presents difficulties. However, we are able to show that it is negligible under regularity conditions. The results are in a good form when $y \in C$, but not so good when only $y \in D$. (Recall that the limit z in (1.16) may be discontinuous even if $y \in C$, so the case $y \in C$ is interesting and important.)

We now obtain results about the second term in (2.1) for general functions $\phi : (D_1, d_1) \rightarrow (D_2, d_2)$, where $D_i \equiv D([0, t_i], \mathbb{R}^{k_i})$ for $i = 1, 2$.

Theorem 9.2.1. (reduction of convergence preservation to the derivative)
Suppose that $\phi : (D_1, d_1) \rightarrow (D_2, d_2)$, where the metrics d_i satisfy the properties:

$$d_i(cx_1, cx_2) = cd_i(x_1, x_2) \quad \text{for all } c > 0, i = 1, 2, \quad (2.2)$$

$$d_i(x_1 + x_3, x_2 + x_3) = d_i(x_1, x_2), \quad i = 1, 2, \quad (2.3)$$

$$d_2(\phi(x_1), \phi(x_2)) \leq Kd_1(x_1, x_2) \quad \text{for some } K > 0, \quad (2.4)$$

for all $x_1, x_2, x_3 \in D_i$. Then

$$d_2(\epsilon^{-1}[\phi(x_\epsilon) - \phi(x)], \epsilon^{-1}[\phi(x + \epsilon y) - \phi(x)]) \leq Kd_1(\epsilon^{-1}(x_\epsilon - x), y). \quad (2.5)$$

Proof. The conditions imply that

$$\begin{aligned} d_2(\epsilon^{-1}[\phi(x_\epsilon) - \phi(x)], \epsilon^{-1}[\phi(x + \epsilon y) - \phi(x)]) &= \epsilon^{-1}d_2(\phi(x_\epsilon), \phi(x + \epsilon y)) \\ &\leq \epsilon^{-1}Kd_1(x_\epsilon, x + \epsilon y) \\ &= Kd_1(\epsilon^{-1}(x_\epsilon - x), y). \quad \blacksquare \end{aligned}$$

Notice that the uniform metric satisfies conditions (2.2) and (2.3). The following application of Theorem 9.2.1 is elementary.

Theorem 9.2.2. (reduction for the supremum and reflection maps with the uniform metric) If d_1 and d_2 in Theorem 9.2.1 above are the uniform

metric on $D([0, t], \mathbb{R})$ and ϕ is the supremum function in equation (13.4.1) in the book or the reflection map in equation (13.5.1) in the book, then the conditions of Theorem 9.2.1 above are satisfied, so that conclusion (2.5) holds.

Proof. It is evident that the uniform metric on D satisfies conditions (2.2) and (2.3). The supremum and reflection functions also satisfy (2.4) with respect to the uniform metric by Lemmas 13.4.1 and 13.5.1 in the book.

Example 9.2.1. *The need for the map ϕ to be Lipschitz.* To see the need for $\phi : D \rightarrow D$ being Lipschitz in Theorem 9.2.1, let $\phi(x)(t) = \sqrt{x(1)}$, $t \geq 0$. If $\|x_\epsilon - x\|_t \rightarrow 0$ for $t > 1$, then $\|\phi(x_\epsilon) - \phi(x)\|_t \rightarrow 0$, but ϕ is not Lipschitz. Suppose that $x(t) = 0$, $y(t) = 1$ and $x_\epsilon(t) = x(t) + \epsilon y(t) = \epsilon$, $t \geq 0$. Then $\|\epsilon^{-1}(x_\epsilon - x) - y\| = 0$ for all ϵ ,

$$\epsilon^{-1}[\phi(x_\epsilon) - \phi(x)](t) = \epsilon^{-1}[\sqrt{\epsilon} - 0] = \epsilon^{-1/2} \rightarrow \infty \quad \text{as } \epsilon \downarrow 0. \quad \blacksquare \quad (2.6)$$

Unfortunately, for the non-uniform Skorohod metrics on D , which we will want to consider when $y \notin C$, we do not have properties (2.2) and (2.3) in Theorem 9.2.1.

Example 9.2.2. *Failure for nonuniform metrics.* Unlike with the uniform metric, we cannot conclude that $d(\epsilon^{-1}x_\epsilon, \epsilon^{-1}x + y) \rightarrow 0$ as $\epsilon \downarrow 0$ when $d(\epsilon^{-1}(x_\epsilon - x), y) \rightarrow 0$ as $\epsilon \downarrow 0$ if d is the J_1 , M_1 or M_2 metric and y is not continuous. To see this, let $x(t) = tI_{[0,1)}(t) + (2-t)I_{[1,2]}(t)$, $y = I_{[0,1)} - I_{[1,2]}$ and $x_\epsilon = (x + \epsilon)I_{[0,1-\epsilon]} + (x - \epsilon)I_{[1-\epsilon,2]}$ in $D([0, 2], \mathbb{R})$. Then $\epsilon^{-1}(x_\epsilon - x) = y \circ \lambda_\epsilon$, where $\lambda_\epsilon \in \Lambda$ with $\lambda_\epsilon(1) = 1 - \epsilon$, $\lambda_\epsilon(0) = 0$ and $\lambda_\epsilon(2) = 2$. Hence $d_{J_1}(\epsilon^{-1}(x_\epsilon - x), y) = \|\lambda_\epsilon - e\| = \epsilon \rightarrow 0$ as $\epsilon \downarrow 0$. However $\epsilon^{-1}x_\epsilon^\uparrow(2) = \epsilon^{-1}x_\epsilon^\uparrow(1 - \epsilon) = \epsilon^{-1}$, while $(\epsilon^{-1}x + y)^\uparrow(2) = (\epsilon^{-1}x + y)(1-) = \epsilon^{-1} + 1$, so that $d_{M_2}(\epsilon^{-1}x_\epsilon, \epsilon^{-1}x + y) \geq 1$. $\blacksquare$

However, under regularity conditions, we can also establish results starting from J_1 , M_1 and M_2 convergence. We state the following results for the strong SJ_1 , SM_1 and SM_2 metrics on $D([0, t], \mathbb{R}^k)$. Corresponding results for the product metrics for Lemmas 9.2.1 and 9.2.2 below follow; just consider one coordinate at a time.

Recall that x is Lipschitz on $[0, t]$ if there is a constant K so that $|x(t_1) - x(t_2)| \leq K|t_1 - t_2|$ for $0 \leq t_1, t_2 \leq t$. This regularity condition is typically satisfied in applications, because x often satisfies an ordinary differential equation (ODE). If x is absolutely continuous with derivative $\dot{x}$, where $\dot{x} \in$

D , then for each $t > 0$, there exists K such that $|\dot{x}(s)| \leq K$ for $0 \leq s \leq t$ and, for $0 \leq t_1 < t_2 \leq t$,

$$|x(t_2) - x(t_1)| \leq \int_{t_1}^{t_2} |\dot{x}(s)| ds \leq K|t_2 - t_1|, \quad (2.7)$$

so that x is Lipschitz.

Lemma 9.2.1. (subtracting a common Lipschitz function) *Suppose that x is Lipschitz in $[0, t]$ with Lipschitz constant K . If d_t is the SJ_1 , SM_1 or SM_2 metric on $D([0, t], \mathbb{R}^k)$, then*

$$d_t(x_1 - x, x_2 - x) \leq (1 + K)d_t(x_1, x_2). \quad (2.8)$$

Proof. First consider J_1 . For all $\epsilon > 0$, there exist $\eta(\epsilon) > 0$ and increasing homeomorphisms λ_ϵ of $[0, t]$ such that

$$\|x_1 - x_2 \circ \lambda_\epsilon\|_t \vee \|\lambda_\epsilon - e\|_t \leq (1 + \eta(\epsilon))d_t(x_1, x_2).$$

It follows that

$$\begin{aligned} \|(x_1 - x) - (x_2 - x) \circ \lambda_\epsilon\|_t &\leq \|x_1 - x_2 \circ \lambda_\epsilon\|_t + \|x - x \circ \lambda_\epsilon\|_t \\ &\leq (1 + \eta(\epsilon))d_t(x_1, x_2) + K\|\lambda_\epsilon - e\|_t \\ &\leq (1 + \eta(\epsilon) + K[1 + \eta(\epsilon)])d_t(x_1, x_2). \end{aligned}$$

Since $\eta(\epsilon)$ can be made arbitrarily small, the proof for J_1 is complete. Now consider M_1 . For all $\epsilon > 0$ and $t > 0$, there exist $\eta(\epsilon) > 0$ and parametric representations $(u_1, r_{1\epsilon})$ of x_1 and $(u_{2\epsilon}, r_{1\epsilon})$ of x_2 such that

$$\|u_{1\epsilon} - u_{2\epsilon}\| \vee \|r_{1\epsilon} - r_{2\epsilon}\| \leq (1 + \eta(\epsilon))d_t(x_1, x_2).$$

Since x is continuous, $(x \circ r_{1\epsilon}, r_{1\epsilon})$ and $(x \circ r_{2\epsilon}, r_{2\epsilon})$ are parametric representations of x , $(u_{1\epsilon} - x \circ r_{1\epsilon}, r_{1\epsilon})$ and $(u_{2\epsilon} - x \circ r_{2\epsilon}, r_{2\epsilon})$ are parametric representations of $x_1 - x$ and $x_2 - x$, and

$$\begin{aligned} \|(u_{1\epsilon} - x \circ r_{1\epsilon}) - (u_{2\epsilon} - x \circ r_{2\epsilon})\| &\leq \|u_{1\epsilon} - u_{2\epsilon}\| + \|x \circ r_{1\epsilon} - x \circ r_{2\epsilon}\| \\ &\leq (1 + \eta(\epsilon))d_t(x_1, x_2) + K\|r_{1\epsilon} - r_{2\epsilon}\| \\ &\leq (1 + \eta(\epsilon) + K[1 + \eta(\epsilon)])d_t(x_1, x_2). \end{aligned}$$

Since $\eta(\epsilon)$ can be arbitrarily small, the proof for M_1 is complete. Now consider M_2 . let $(z_1, t_1) \in \Gamma_{x_1}$. If $(z_2, t_2) \in \Gamma_{x_2}$ is such that $\|(z_1, t_1) - (z_2, t_2)\| < \delta$, then $(z_1 - x(t_1), t_1) \in \Gamma_{x_1 - x}$, $(z_2 - x(t_2), t_2) \in \Gamma_{x_2 - x}$ and

$$\begin{aligned} \|(z_1 - x(t_1), t_1) \vee (z_2 - x(t_2), t_2)\| &\leq \|(z_1, t_1) - (z_2, t_2)\| + \|x(t_1) - x(t_2)\| \\ &\leq \delta + K\|t_1 - t_2\| \leq (1 + K)\delta. \quad \blacksquare \end{aligned}$$

Next we generalize (2.2).

Lemma 9.2.2. (deterministic scaling) *Let d_t be the SJ_1 , SM_1 or SM_2 metric on $D([0, t], \mathbb{R}^k)$. For any $c > 0$,*

$$d_{ct}(cx_1 \circ c^{-1}e, cx_2 \circ c^{-1}e) = cd_t(x_1, x_2) \quad (2.9)$$

or, equivalently,

$$d_t(cx_1, cx_2) = cd_{t/c}(x_1 \circ ce, x_2 \circ ce) \leq (c \vee 1)d_t(x_1, x_2) . \quad (2.10)$$

Proof. First, for SJ_1 , note that $\lambda \in \Lambda_t$ if and only if $c\lambda \circ c^{-1}e \in \Lambda_{ct}$ for $c > 0$ and

$$\|c\lambda \circ c^{-1}e - e\|_{ct} = c\|\lambda - e\|_t .$$

Hence

$$\begin{aligned} & d_{ct}(cx_1 \circ c^{-1}e, cx_2 \circ c^{-1}e) \\ &= \inf_{\lambda \in \Lambda_t} \{ \|cx_1 \circ c^{-1}e - (cx_2 \circ c^{-1}e) \circ (c\lambda \circ c^{-1}e)\|_{ct} \vee \|c\lambda \circ c^{-1}e - e\|_{ct} \} \\ &= \inf_{\lambda \in \Lambda_t} \{ c\|x_1 - x_2 \circ \lambda\|_t \vee c\|\lambda - e\|_t \} \\ &= cd_t(x_1, x_2) . \end{aligned}$$

Next, for SM_2 , note that $c\Gamma_{x_i}$ is the graph of $cx_i \circ c^{-1}e$ over $[0, ct]$ if and only if Γ_{x_i} is the graph of x_i over $[0, t]$. Hence (2.9) holds. Finally, for SM_1 , note that (cu_i, cr_i) is a parametric representation of $cx_i \circ c^{-1}e$ over $[0, ct]$ if and only if (u_i, r_i) is a parametric representation of x_i over $[0, t]$. Hence (2.9) holds. ■

Our next result goes beyond Theorem 9.2.1 by allowing the map ϕ to be Lipschitz with respect to the SJ_1 , SM_1 or SM_2 metrics, but not the uniform metric.

Theorem 9.2.3. (Lipschitz functions with respect to non-uniform metrics) *Suppose that $y \in D([0, t_1], \mathbb{R}^{k_1})$ and $x, x_\epsilon, x + \epsilon y$ all belong to a subset A of $D([0, t_1], \mathbb{R}^{k_1})$ for sufficiently small $\epsilon > 0$. Suppose that $\phi : A \rightarrow D([0, t_2], \mathbb{R}^{k_2})$ is Lipschitz with respect to the metrics d_1 on A and d_2 on $D([0, t_2], \mathbb{R}^{k_2})$, i.e., there is a constant K such that*

$$d_2(\phi(x_1), \phi(x_2)) \leq Kd_1(x_1, x_2) \quad (2.11)$$

for all $x_1, x_2 \in A$, where d_1 and d_2 are non-uniform Skorohod metrics (not necessarily the same). Suppose that x is Lipschitz on $[0, t_1]$ and $\phi(x)$ is Lipschitz on $[0, t_2]$. Then there is a constant K' such that

$$\begin{aligned} & d_2(\epsilon^{-1}[\phi(x_\epsilon) - \phi(x)], \epsilon^{-1}[\phi(x + \epsilon y) - \phi(x)]) \\ & \leq K'\epsilon^{-1}d_1(x_\epsilon - x, \epsilon y) \\ & \leq K'\|\epsilon^{-1}(x_\epsilon - x) - y\|_{t_1} . \end{aligned} \quad (2.12)$$

Proof. By Lemmas 9.2.2 and 9.2.1 and the assumptions, for $\epsilon < 1$, there are constants K_1 , K_2 and K_3 such that

$$\begin{aligned}
& d_2(\epsilon^{-1}[\phi(x_\epsilon) - \phi(x)], \epsilon^{-1}[\phi(x + \epsilon y) - \phi(x)]) \\
& \leq \epsilon^{-1} d_2(\phi(x_\epsilon) - \phi(x), \phi(x + \epsilon y) - \phi(x)) \\
& \leq K_1 \epsilon^{-1} d_2(\phi(x_\epsilon), \phi(x + \epsilon y)) \\
& \leq K_1 K_2 \epsilon^{-1} d_1(x_\epsilon, x + \epsilon y) \\
& \leq K_1 K_2 K_3 \epsilon^{-1} d_1(x_\epsilon - x, \epsilon y) \\
& \leq K_1 K_2 K_3 \epsilon^{-1} \|x_\epsilon - x - \epsilon y\|_{t_1} \\
& \leq K_1 K_2 K_3 \|\epsilon^{-1}(x_\epsilon - x) - y\|_{t_1} . \quad \blacksquare
\end{aligned} \tag{2.13}$$

The final upper bound in Theorem 9.2.3 does not help with the supremum and reflection maps because the supremum and reflection maps already have the required Lipschitz properties with respect to the uniform metric, by Theorem 9.2.2. In order to apply Theorem 9.2.3 without having to resort to the cruder uniform metric bound, we need to have

$$d_1(x_\epsilon - x, \epsilon y) = o(\epsilon) \quad \text{as } \epsilon \downarrow 0. \tag{2.14}$$

First, from this analysis, we see the need to be precise about what we mean about $o(\epsilon)$ terms in (1.12)–(1.15). Next, we observe that $d_1(\epsilon^{-1}[x_\epsilon - x], y) \rightarrow 0$ does not directly imply that $d_1(x_\epsilon - x, \epsilon y) = o(\epsilon)$ as $\epsilon \downarrow 0$, but that it is possible to have $d_1(x_\epsilon - x, \epsilon y) = o(\epsilon)$ as $\epsilon \downarrow 0$ without having $\|x_\epsilon - x - \epsilon y\|_{t_1} = o(\epsilon)$ as $\epsilon \downarrow 0$.

Example 9.2.3. *Condition (2.14) is weaker than the usual limit.* We would like to have $\epsilon^{-1}d_t(x_\epsilon - x, \epsilon y) \rightarrow 0$ as $\epsilon \downarrow 0$ whenever $d_t(\epsilon^{-1}(x_\epsilon - x), y) \rightarrow 0$ as $\epsilon \downarrow 0$, so that we could improve upon (2.12), but that implication is not valid. To see that, let $x = y = I_{[1,2]}$ in $D([0, 2], \mathbb{R})$ and let $x_\epsilon = x + \epsilon I_{[1+\delta_\epsilon, 2]}$. Then $\epsilon^{-1}(x_\epsilon - x) = I_{[1+\delta_\epsilon, 2]}$ and $d_{J_1}(\epsilon^{-1}(x_\epsilon - x), y) = \delta_\epsilon$. On the other hand $\epsilon^{-1}d_{J_1}(x_\epsilon - x, \epsilon y) = \epsilon^{-1}(\epsilon \wedge \delta_\epsilon)$, which converges to 0 if and only if $\epsilon^{-1}\delta_\epsilon \rightarrow 0$ as $\epsilon \rightarrow 0$. Hence, we do not necessarily have $\epsilon^{-1}d_t(x_\epsilon - x, \epsilon y) \rightarrow 0$ as $\epsilon \downarrow 0$, given $d_t(\epsilon^{-1}(x_\epsilon - x), y) \rightarrow 0$, but we could have it, as is the case here when $\epsilon^{-1}\delta_\epsilon \rightarrow 0$ as $\epsilon \downarrow 0$. On the other hand, $\|\epsilon^{-1}(x_\epsilon - x) - y\| = 1$ for all $\epsilon > 0$. $\blacksquare$

Example 9.2.4. *A parametric family of examples.* Consider Example 9.2.2 modified by having

$$x_\epsilon = (x + \epsilon)I_{[0, 1-\epsilon^p]} + (x - \epsilon)I_{[1-\epsilon^p, 2]}. \tag{2.15}$$

Then $x_\epsilon - x = \epsilon y \circ \lambda_\epsilon$ for $\lambda_\epsilon(1) = 1 - \epsilon^p$, $\lambda_\epsilon(0) = 0$ and $\lambda_\epsilon(2) = 2$ with λ_ϵ defined by linear interpolation elsewhere. Thus

$$d_{J_1}(x_\epsilon - x, \epsilon y) = d_{J_1}(\epsilon^{-1}(x_\epsilon - x), y) = \|\lambda_\epsilon - e\| = \epsilon^p, \quad (2.16)$$

so that condition (2.14) holds if $p > 1$, but not if $0 < p \leq 1$.

9.3. Derivative of the Supremum Function

In this section we consider the derivative of the supremum function; i.e., we find conditions under which the limit (1.15) is valid and identify the limit z when $\phi : D \rightarrow D$ is the supremum function. The supremum function maps $x \in D \equiv D([0, T], \mathbb{R})$ into $x^\uparrow \in D$ for

$$x^\uparrow(t) \equiv \sup_{0 \leq s \leq t} x(s), \quad 0 \leq t \leq T. \quad (3.1)$$

In order to treat the derivatives, we will find it necessary to consider functions outside of D . Thus let D_{lim} be the set of functions with left and right limits everywhere, but without having to be either left continuous or right continuous at each discontinuity point. In general, we will only be able to conclude (in Theorem 9.3.2 below) that the derivative belongs to D_{lim} . In our definition of the derivative, we start by allowing one function to be in D_{lim} . For $x \in D$, $y \in D_{lim}$ and $\epsilon > 0$, let

$$z_\epsilon \equiv z_\epsilon(x, y) \equiv \epsilon^{-1}[(x + \epsilon y)^\uparrow - x^\uparrow] = (\epsilon^{-1}x + y)^\uparrow - \epsilon^{-1}x^\uparrow. \quad (3.2)$$

The derivative of the supremum function (in the direction y , evaluated at x) is the limit of z_ϵ as $\epsilon \downarrow 0$, if it exists. We will show that the limit does exist under regularity conditions and identify it. In this section we consider pointwise convergence for all t ; in the next section we consider M_1 and M_2 convergence.

We start by stating two elementary lemmas; the second follows from the first.

Lemma 9.3.1. (the case of constant y) *If $y(s) = c$, $0 \leq s \leq t$, then $z_\epsilon(t) = c$ for all ϵ .*

For $z^\downarrow$ be the infimum function; i.e., $z^\downarrow = -(-z)^\uparrow$.

Lemma 9.3.2. (monotone bounds) *For all $\epsilon > 0$, $y^\downarrow \leq z_\epsilon \leq y^\uparrow$.*

Even though x is right-continuous, it can approach its supremum from the left ($x(s) = sI_{[0,t_0)}(s)$) or right ($x(s) = -sI_{[t_0,t_1)}(s)$). Let $\Phi_x^L(t)$ and $\Phi_x^R(t)$ be the subsets of time points in $[0, t]$ at which the left and right limits of x attain the supremum; i.e.,

$$\Phi_x^L(t) = \{s : 0 < s \leq t, x(s-) = x^\uparrow(t)\} \quad (3.3)$$

and

$$\Phi_x^R(t) = \{s : 0 \leq s \leq t, x(s+) = x^\uparrow(t)\}. \quad (3.4)$$

Let $\Phi_x(t) = \Phi_x^L(t) \cup \Phi_x^R(t)$. When $x \in C$, $\Phi_x^L(t) = \Phi_x^R(t)$.

Example 9.3.1. *The possibility of empty sets.* It is possible for Φ_x^L or Φ_x^R to be empty: Let $x(t) = tI_{[0,1)}(t)$, $t \geq 0$. Then, for $t \geq 1$, $\Phi_x^L(t) = \{1\}$, while $\Phi_x^R(t) = \emptyset$. However, $\Phi_x^L(t) \cup \Phi_x^R(t) \neq \emptyset$. ■

These subsets need not be closed, but they have the following partial closure property.

Lemma 9.3.3. (partial closure property) *For any $x \in D$ and $t \geq 0$, $\Phi_x^L(t)$ is closed from the left, while $\Phi_x^R(t)$ is closed from the right; i.e., if $s_n \uparrow s$ in $[0, t]$ and $s_n \in \Phi_x^L(t)$ for all n , then $s \in \Phi_x^L(t)$; if $s_n \downarrow s$ and $s_n \in \Phi_x^R(t)$ for all n , then $s \in \Phi_x^R(t)$. Moreover, if $s_n \uparrow s$ in $[0, t]$ and $s_n \in \Phi_x^R(t)$ for all n , then $s \in \Phi_x^L(t)$; if $s_n \downarrow s$ in $[0, t]$ and $s_n \in \Phi_x^L(t)$ for all n , then $s \in \Phi_x^R(t)$.*

Corollary 9.3.1. (compactness of $\Phi_x(t)$) *For each $t > 0$, $\Phi_x(t)$ is a compact subset of $[0, t]$.*

We next show that z_ϵ is monotone in ϵ .

Lemma 9.3.4. (monotonicity in ϵ) *For z_ϵ in (3.2), $z_\epsilon(t)$ decreases as ϵ decreases for each t .*

Proof. We want to show that

$$(\epsilon_2^{-1}x + y)^\uparrow - \epsilon_2^{-1}x^\uparrow < (\epsilon_1^{-1}x + y)^\uparrow - \epsilon_1^{-1}x^\uparrow$$

for $\epsilon_1 > \epsilon_2$ or, equivalently,

$$(\epsilon_2^{-1}x + y)^\uparrow - (\epsilon_1^{-1}x + y)^\uparrow < (\epsilon_2^{-1} - \epsilon_1^{-1})x^\uparrow. \quad (3.5)$$

However, (3.5) follows from the relation

$$x_1^\uparrow - x_2^\uparrow \leq (x_1 - x_2)^\uparrow. \quad \blacksquare$$

We first establish pointwise convergence for z_ϵ in (3.2).

Theorem 9.3.1. (pointwise convergence) *For each $x \in D, y \in D_{lim}$ and $t \geq 0$,*

$$\lim_{\epsilon \downarrow 0} z_\epsilon(t) = z(t) \equiv \sup_{s \in \Phi_x^L(t)} y(s-) \vee \sup_{s \in \Phi_x^R(t)} \{y(s), y(s+)\} . \quad (3.6)$$

Proof. The convergence follows from the monotonicity established in Lemma 9.3.3. Lemma 9.3.2 above provides a lower bound, which implies that there is a proper limit for each t . For any $\delta > 0$, let $s_\epsilon(t)$ be a point in $[0, t]$ such that

$$(\epsilon^{-1}x + y)(s_\epsilon(t)) \geq (\epsilon^{-1}x + y)^\uparrow(t) - \delta . \quad (3.7)$$

(Since x and y need not be continuous, the supremum of $\epsilon^{-1}x + y$ need not be attained.) Then

$$\begin{aligned} y(s_\epsilon(t)) &\geq y(s_\epsilon(t)) + \epsilon^{-1}\{x[s_\epsilon(t)] - x^\uparrow(t)\} \\ &\geq y(s) + \epsilon^{-1}[x(s) - x^\uparrow(t)] - \delta \quad \text{for } 0 \leq s \leq t \\ &\geq \begin{cases} y(s-) - \delta & \text{for } s \in \Phi_x^L(t) \\ y(s) - \delta & \text{for } s \in \Phi_x^R(t) \\ y(s+) - \delta & \text{for } s \in \Phi_x^R(t) , \end{cases} \end{aligned} \quad (3.8)$$

implying that

$$\varliminf_{n \rightarrow \infty} y(s_\epsilon(t)) \geq z(t), \quad t \geq 0 . \quad (3.9)$$

We now verify that

$$\varlimsup_{n \rightarrow \infty} y(s_\epsilon(t)) \leq z(t), \quad t \geq 0 . \quad (3.10)$$

Start by choosing $\{s_\epsilon(t)\}$ such that $y(s_\epsilon(t)) \rightarrow \varlimsup_{n \rightarrow \infty} y(s_\epsilon(t))$ as $\epsilon \downarrow 0$. Since

$s_\epsilon(t) \in [0, t]$, any subsequence from $\{s_\epsilon(t)\}$ has a convergent subsequence $\{s_{\epsilon'}(t)\}$ as $\epsilon' \downarrow 0$. (Let $\epsilon' \downarrow 0$ through countably many values.) So suppose that $s_{\epsilon'}(t) \rightarrow s_0(t)$ as $\epsilon' \downarrow 0$. Without loss of generality, by taking a further subsequence if necessary, we can assume that either $s_{\epsilon'}(t) \uparrow s_0(t)$ with $s_{\epsilon'}(t) < s_0(t)$ for all $\epsilon' > 0$ or $s_{\epsilon'}(t) \downarrow s_0(t)$ with $s_{\epsilon'}(t) \geq s_0(t)$ for all $\epsilon' > 0$. Suppose that $s_{\epsilon'}(t) \uparrow s_0(t)$. Then $y(s_{\epsilon'}(t)) \rightarrow y(s_0(t)-)$. We can deduce from (3.8) that there is a constant K such that, for all ϵ' ,

$$-K \leq \epsilon^{-1}[x(s_{\epsilon'}(t)) - x^\uparrow(t)] \leq 0 , \quad (3.11)$$

implying that $x(s_{\epsilon'}(t)) \rightarrow x^\uparrow(t)$ as $\epsilon' \rightarrow 0$, so that $x(s_0(t)-) = x^\uparrow(t)$ and $s_0(t) \in \Phi_x^L(t)$. By this argument,

$$\overline{\lim}_{n \rightarrow \infty} y(s_\epsilon(t)) \leq \sup_{s \in \Phi_x^L(t)} y(s-) . \quad (3.12)$$

On the other hand, if $s_{\epsilon'}(t) \downarrow s_0(t)$, we can deduce by the same reasoning that

$$\overline{\lim}_{n \rightarrow \infty} y(s_\epsilon(t)) \leq \sup_{s \in \Phi_x^R(t)} \{y(s), y(s+)\} . \quad (3.13)$$

Since one of (3.12) or (3.13) must hold, we have established (3.10). Finally, from the first and last lines of (3.8),

$$0 \geq \epsilon^{-1} \{x(s_\epsilon(t)) - x^\uparrow(t)\} \geq z(t) - y(s_\epsilon(t)) . \quad (3.14)$$

Since $y(s_\epsilon(t)) \rightarrow z(t)$, $\epsilon^{-1} \{x(s_\epsilon(t)) - x^\uparrow(t)\} \rightarrow 0$ as $\epsilon \downarrow 0$, which implies that $z_\epsilon(t) \rightarrow z(t)$ as $\epsilon \downarrow 0$. ■

Corollary 9.3.2. (simplification under extra conditions) *Suppose that $x \in C$ and $y \in D_{lim}$. Then the limit z in (3.6) is*

$$z(t) = \sup_{s \in \Phi_x(t)} \{y(s-), y(s), y(s+)\} . \quad (3.15)$$

If, in addition, $y \in C$, then

$$z(t) = \sup_{s \in \Phi_x(t)} \{y(s)\} . \quad (3.16)$$

We now determine the structure of the limit function z in (3.6). Since $\Phi_x^L(t)$, $\Phi_x^R(t)$ and $\Phi_x(t)$ are subsets of $[0, t]$, we need a notion of convergence of sets. For subsets A_n and A of $\mathbb{R}$, we say that $A_n \rightarrow A$ if (i) for all $a_n \in A_n$, $n \geq 1$, $\{a_n\}$ has a convergent subsequence and the limits of all convergent subsequences belong to A , and (ii) for all $a \in A$, there exists $a_n \in A_n$, $n \geq 1$, such that $a_n \rightarrow a$ as $n \rightarrow \infty$. In our set limits involving $\Phi_x^L(t)$ and $\Phi_x^R(t)$, only three special cases arise: (i) A_n is independent of n for all sufficiently large n , (ii) the sequence $\{A_n\}$ is eventually monotone, i.e., either $A_n \subseteq A_{n+1}$ for all sufficiently large n or $A_n \supset A_{n+1}$ for all sufficiently large n , and (iii) $A = \{a\}$, i.e., the limit set contains a single point.

When we consider $\Phi_x(t) \equiv \Phi_x^L(t) \cup \Phi_x^R(t)$, we have compact subsets of $[0, t]$. Then the notion of set convergence above is induced by the Hausdorff metric on the space $\mathcal{C} \equiv \mathcal{C}([0, \infty))$ of compact subsets of $[0, \infty)$, defined in (2.8) in Chapter V.

However, even if x and $x^\uparrow$ are continuous in t , $\Phi_x(t)$ is in general *not* continuous in t . Moreover, at some time points, $\Phi_x(t)$ is neither left-continuous nor right-continuous.

Example 9.3.2. *Lack of continuity from left or right in $\Phi_x(t)$.* Suppose that $x(t) = (1-t)I_{[0,1)}(t) + (t-1)I_{[1,\infty)}(t)$. Then $\Phi_x(t) = \{0\}$, $0 \leq t < 2$, $\Phi_x(2) = \{0, 1\}$ and $\Phi_x(t-1) = \{t\}$, $t > 2$, so that Φ_x is neither left-continuous nor right-continuous at $t = 2$. However, $\Phi_x(2)$ is the union of the left and right limits $\Phi_x(2-)$ and $\Phi_x(2+)$. ■

Example 9.3.3. *Neither left-continuous everywhere nor right-continuous everywhere.* We can extend Example 9.3.2 to show that the limit z need not be either a left-continuous function or a right-continuous function, even if x and y are both continuous. Let

$$x(t) = (1-t)I_{[0,1)}(t) + (t-1)I_{[1,3)}(t) + (5-t)I_{[3,4)}(t) + (t-3)I_{[4,\infty)}(t) \quad (3.17)$$

and

$$y(t) = -tI_{[0,2.5]} + 6(t-2.5)I_{[2.5,\infty)}(t). \quad (3.18)$$

Then

$$\begin{aligned} \Phi_x(t) &= \{0\}, & 0 \leq t < 2, & \quad \Phi_x(2) = \{0, 2\} \\ \Phi_x(t) &= \{t\}, & 2 < t \leq 3, & \quad \Phi_x(t) = \{3\}, \quad 3 \leq t < 5, \\ \Phi_x(5) &= \{3, 5\}, & \Phi_x(t) = \{t\}, & \quad t > 5, \\ z(2) &= 0 \quad \text{and} \quad z(5) = 15. \end{aligned} \quad (3.19)$$

Then z is discontinuous at $t = 2$ and $t = 5$, with z being left-continuous at 2 and right-continuous at 5. Hence z is neither left-continuous everywhere nor right-continuous everywhere. On the positive side, z is either left-continuous or right-continuous at each t and z is upper semicontinuous everywhere. ■

Example 9.3.4. *Neither left-continuous nor right-continuous at one t .* We now show that the limit z in (3.8) need not be either left-continuous or right-continuous at a single argument t when $x \in C$ and $y \in D$ but $y \notin C$. We construct y and x so that y and Φ_x have only one common discontinuity. Let

$$y(t) = tI_{[0,1)}(t) + I_{[1,2)}(t), \quad t \geq 0, \quad (3.20)$$

and

$$x(t) = -tI_{[0,1)}(t) + (t-2)I_{[1,\infty)}(t), \quad t \geq 0, \quad (3.21)$$

so that

$$\Phi_x(t) = \{0\}, 0 \leq t < 2, \Phi_x(2) = \{0, 2\} \quad \text{and} \quad \Phi_x(t) = t, \quad t > 2. \quad (3.22)$$

Hence y and Φ_x are continuous everywhere except $t = 2$. Moreover,

$$z(2) = \sup_{s \in \{0, 2\}} \{y(s)\} \vee \sup_{s \in \{2\}} \{y(s-)\} = 0 \vee 1 = 1, \quad (3.23)$$

while $z(t) = 0$ for all other t . Hence the left and right limits coincide at $t = 2$ but do not equal $z(2)$, so that $z \notin D$. It is easy to see that $z_\epsilon(2) = 1$ and

$$z_\epsilon(t) = 0, \quad 0 \leq t \leq 2 - \epsilon \quad \text{and} \quad t \geq 2 + \epsilon,$$

with z_ϵ defined by linear interpolation elsewhere. Hence, z_ϵ has slope ϵ^{-1} on $[2 - \epsilon, 2]$, slope $-\epsilon^{-1}$ on $[2, 2 + \epsilon]$ and is 0 elsewhere. Consistent with Theorem 9.3.1, z_ϵ converges pointwise to z . We will want to impose regularity conditions to prevent such pathological behavior. As an alternative, we could conclude that z_ϵ converges to a limit in one of the larger spaces E or F in Chapter X. ■

We now introduce a regularity condition under which the limit z in (3.6) has left and right limits everywhere and is either left continuous or right continuous everywhere (without necessarily being right continuous everywhere). Let $D_{l,r}$ denote this space. We first define some subsets of $[0, \infty)$. (We could alternatively restrict attention to a subinterval $[0, T]$.) For any $x \in D$, let $Rinc(x)$ and $Linc(x)$ be the set of right-increase and left-increase points of x , let $Lconst(x)$ be the set of left-constant points of x , and let $Amax(x)$ be the argmax set of x , i.e., the set of arguments at which x equals its supremum, i.e.,

$$\begin{aligned} Rinc(x) &\equiv \{t \geq 0 : x(t) < x(t + \epsilon) \text{ for all sufficiently small } \epsilon\} \\ Linc(x) &\equiv \{t \geq 0 : x(t - \epsilon) < x(t) \text{ for all sufficiently small } \epsilon\} \\ Lconst(x) &\equiv \{t \geq 0 : x(t - \epsilon) = x(t) \text{ for all sufficiently small } \epsilon\} \\ Amax(x) &\equiv \{t \geq 0 : t \in \Phi_x^R(t)\}. \end{aligned} \quad (3.27)$$

We will look at these sets for the functions x and $x^\uparrow$. Of course, $x^\uparrow$ is nondecreasing and right-continuous. Let $Disc(x)$ be the set of discontinuity points of x .

Theorem 9.3.2. (regularity properties of the limit z) *Suppose that $x, y \in D$. Then $z \in D_{lim}$, where z is the limit in (3.6). At all t not in the set*

$$Bad(x) \equiv Rinc(x^\uparrow) \cap Lconst(x^\uparrow) \cap Disc(x)^c \cap Linc(x) \cap Amax(x), \quad (3.28)$$

z is either left-continuous or right-continuous. For $t \in \text{Bad}(x)$, $z(t+) = y(t)$, $z(t-)$ is independent of $\{y(t-), y(t)\}$ and $z(t) = \max\{z(t-), y(t-), y(t)\}$, so that z is left-continuous at t if $z(t-) \geq y(t-) \vee y(t)$, right-continuous at t if $y(t) \geq y(t-) \vee z(t-)$, and neither left-continuous nor right-continuous if $y(t-) > y(t) \vee z(t-)$. If

$$y(t-) \leq z(t-) \vee y(t) \quad (3.29)$$

for all $t \in \text{Bad}(x)$, for which a sufficient condition is

$$\text{Disc}(y) \cap \text{Bad}(x) = \emptyset, \quad (3.30)$$

then z is either left-continuous or right-continuous at all t , so that $z \in D_{l,r}$.

Corollary 9.3.3. (regularity for continuous y) If $x \in D$ and $y \in C$, then $z \in D_{l,r}$.

Remark 9.3.1. Sufficient condition for having more than one point in the set. Let $|\Phi_x(t)|$ be the cardinality of the set $\Phi_x(t)$. Note that $|\Phi_x(t)| \geq 2$ when $t \in \text{Lconst}(x^\uparrow) \cap \text{Amax}(x)$, i.e.,

$$\text{Lconst}(x^\uparrow) \cap \text{Amax}(x) \subseteq \{t : |\Phi_x(t)| \geq 2\}, \quad (3.31)$$

so that $t \in \text{Bad}(x)$ when $|\Phi_x(t)| \geq 2$ and $x(t - \epsilon) < x(t-) = x(t) = x^\uparrow(t) < x^\uparrow(t + \epsilon)$ for all suitably small $\epsilon > 0$. ■

Remark 9.3.2. The set $\text{Bad}(x)$ is at most countably infinite. From (3.28), it follows that $\text{Bad}(x) \subseteq \text{Disc}(\Phi_x)$, where $\Phi_x \in D([0, \infty), (\mathcal{C}, h))$. Therefore, $\text{Bad}(x)$ is a countable set. ■

Corollary 9.3.4. (regularity properties of the limit Z when Y is a stochastic process) Suppose that $\{Y(t) : t \geq 0\}$ is a stochastic process with sample paths in D . If $x \in D$ and if $P(t \in \text{Disc}(Y)) = 0$ for each $t > 0$, then $P(Z \in D_{l,r}) = 1$, where Z is the limiting stochastic process defined by applying (3.6) to Y .

Proof. In Remark 9.3.2 it was noted that the set $\text{Bad}(x)$ in (3.28) is countable. Consequently,

$$P(Z \in D_{l,r}) = P(\text{Disc}(Y) \cap \text{Bad}(x) = \emptyset) = 1. \quad \blacksquare \quad (3.32)$$

Theorem 9.3.2 is proved by examining all relevant cases. We identify appropriate cases and results for those cases in the following theorem.

Theorem 9.3.3. (identification of relevant cases) *The following is a set of exhaustive and mutually exclusive cases and subcases when $x, y \in D$:*

1. $t \notin Amax(x)$, i.e., $t \notin \Phi_x^R(t)$: z is right-continuous with a left limit at t .
2. $\Phi_x^R(t) = \Phi_x^L(t) = \{t\}$: $z(t) = y(t-) \vee y(t)$, z is either right-continuous or left-continuous at t .
3. $\Phi_x^R(t) = \{t\}$, $\Phi_x^L(t) = \emptyset$: z is right-continuous with a left limit at t .
4. $t \in \Phi_x^R(t) \subseteq \Phi_x(t) \neq \{t\}$, so that cases 1–3 do not hold;
 - (a) $t \notin Rinc(x^\uparrow)$, i.e., $\Phi_x(t) \subseteq \Phi_x(u)$ for some $u > t$: z is right-continuous with a left limit at t .
 - (b) Condition (a) does not hold and $t \in Lconst(x^\uparrow) \cap Linc(x)^c$, i.e., t is not isolated in $\Phi_x(t)$: $z(t-) \geq y(t-)$ and $z(t+) = y(t)$, so that z is left (right) continuous at t if $z(t-) \geq (\leq) y(t)$.
 - (c) Condition (a) does not hold, t is isolated in $\Phi_x(t)$ and $t \in Disc(x)$: $z(t+) = y(t)$ and $z(t) = \max\{z(t-), y(t)\}$, so that z is left (right) continuous if $z(t-) \geq (\leq) y(t)$. (In this case $z(t)$ does not depend upon $y(t-)$.)
 - (d) Condition (a) does not hold, t is isolated in $\Phi_x(t)$ and $t \notin Disc(x)$, i.e., $t \in Bad(x)$ in (3.28): $z(t+) = y(t)$, $z(t-)$ is independent of $\{y(t-), y(t)\}$ and $z(t) = \max\{z(t-), y(t-), y(t)\}$. Hence z is neither left-continuous nor right-continuous at t if and only if $y(t-) > z(t-) \vee y(t)$.

Proof. We prove Theorem 9.3.3 by examining all relevant subcases. We provide a further characterization below, but do not give all details. For this purpose, let

$$\Psi_x^L(t) = \{s : 0 < s \leq t, x(s-) = x^\uparrow(t-)\}, \quad (3.33)$$

$$\Psi_x^R(t) = \{s : 0 \leq s \leq t, x(s) = x^\uparrow(t-)\} \quad (3.34)$$

and $\Psi_x(t) = \Psi_x^L(t) \cup \Psi_x^R(t)$.

Case 1: In this case, $x(t) < x^\uparrow(t)$ and $x^\uparrow(t-) = x^\uparrow(t)$. Since x and $x^\uparrow$ are right-continuous, Φ_x^L and Φ_x^R are constant in $[t, t + \epsilon)$ for all suitably small $\epsilon > 0$, so that z is necessarily right-continuous. We identify three subcases:

(i) If $t \notin \Psi_x^L(t)$, then $x(t-) < x^\uparrow(t-)$, so that Φ_x^L and Φ_x^R are constant in $(t - \epsilon, t + \epsilon)$ for all suitably small $\epsilon > 0$, so that z is constant in the same subinterval. (ii) If $t \in \Psi_x^L(t) = \Phi_x^L(t)$ and $\Phi_x(t) \neq \{t\}$, then x jumps down at time t , so that $x(t-) = x^\uparrow(t-) = x^\uparrow(t) > x(t)$. Since $\Phi_x(t) \neq \{t\}$, $x^\uparrow$ must be constant in $(t - \epsilon, t]$ for all suitably small $\epsilon > 0$ and there must exist $s < t$ such that $x(s) = x^\uparrow(t)$ or $x(s-) = x^\uparrow(t)$. Hence for $s < t' < t$, $\Phi_x^L(t')$ and $\Phi_x^R(t')$ increase as t' increases. Since $t \notin \Phi_x^R(t)$, $\Phi_x^R(t') \uparrow \Phi_x^R(t)$ as $t' \uparrow t$, so that Φ_x^R is continuous at t . Since $\Phi_x^L(t')$ increases as t' increases, $\Phi_x^L(t')$ has a limit as $t' \uparrow t$, but this limit set may be separated from $t \in \Phi_x^L(t)$. Hence, in general z is right-continuous with a left limit at t , with $z(t)$ depending upon $y(t-)$ but not $y(t)$. In this case z is continuous at t if and only if $z(t-) \geq y(t-)$. (iii) If $t \in \Psi_x^L(t) = \Phi_x^L(t)$ and $\Phi_x(t) = \{t\}$, then again x jumps down at time t , $x(t-) = x^\uparrow(t-)$. Since $x^\uparrow$ is increasing from the left at t , there exists a sequence $\{t_n\}$ with $t_n \uparrow t$ as $n \rightarrow \infty$ such that $x(t_n \pm) = x^\uparrow(t_n)$ and $\Phi_x(t_n) = \{t_n\}$. Moreover, for any s with $t_n < s < t$, necessarily $\Phi_x(s) \subseteq [t_n, s]$. Hence, $\Phi_x(s) \rightarrow \Phi_x(t)$ as $s \uparrow t$. This implies that z is continuous at t with $z(t) = y(t-)$. We remark that the case $t \in \Psi_x^L(t)$ but $t \notin \Phi_x^L(t)$ cannot occur because it requires $x(t-) = x^\uparrow(t-) < x^\uparrow(t)$, which implies that x make a jump up to a new maximum at time t , i.e., $t \in \Phi_x^R(t)$, which contradicts our original assumption.

Case 2: $\Phi_x^R(t) = \Phi_x^L(t) = \{t\}$.

In this case $x(t-) = x(t) = x^\uparrow(t)$, so that x is continuous at t . Since $\Phi_x(t) = \{t\}$, $\Phi_x(u) \subseteq [t, u]$ for all $u > t$. Hence $\Phi_x(u) \rightarrow \Phi_x(t) = \{t\}$ as $u \downarrow t$, so that Φ_x is right-continuous and z has a limit from the right with $z(t+) = y(t)$. In this case $x^\uparrow$ is increasing at t , and $\Phi_x(s) \rightarrow \Phi_x^L(t)$ as $s \uparrow t$, so that Φ_x is continuous at t and z has the left limit $z(t-) = y(t-)$. Since $z(t) = y(t) \vee y(t-)$, z is either left-continuous or right-continuous at t ; z is continuous at time t if and only if y is.

Case 3: $\Phi_x^R(t) = \{t\}$ and $\Phi_x^L(t) = \phi$.

In this case $x(t-) \neq x(t) = x^\uparrow(t)$, so that x is discontinuous at t . As in case 2 above, $\Phi_x(s) \rightarrow \Phi_x(t) = \{t\}$ as $s \downarrow t$, so that Φ_x is right-continuous at t and z has the right limit $z(t+) = y(t)$. Since $z(t) = y(t)$, z is right-continuous in this case. We identify three subcases: (i) If $t \notin \Psi_x^L(t)$, then $x(t-) < x^\uparrow(t-) < x^\uparrow(t)$, so that x jumps up to a new maximum at time t and Φ_x^L and Φ_x^R are constant in $(t - \epsilon, t)$ for all suitably small ϵ . Hence Φ_x^L , Φ_x^R and z have limits from the left, but may be discontinuous at t . (ii) If $\Psi_x(t) = \{t\}$, then $x(t-) = x^\uparrow(t-) < x^\uparrow(t)$. As in (ii), x jumps up

to a new maximum at t . Since $\Phi_x(t) = \{t\}$, $x^\uparrow$ is increasing from the left at t . Hence, there exists a sequence $\{t_n\}$ with $t_n \uparrow t$ as $n \rightarrow \infty$ such that $x(t_n \pm) = x^\uparrow(t_n) \uparrow x^\uparrow(t-)$ and $\Phi_x(t_n) = \{t_n\}$. Hence $\Phi_x(s) \subseteq [t_n, s]$ for all s with $t_n < s < t$. Hence, $\Phi_x(s) \rightarrow \Psi_x^L(t) = \{t\}$ as $s \uparrow t$, so that Φ_x and z have limits from the left at t , with $z(t-) = y(t-)$. (iii) Suppose that $\Phi_x^L(t) = \phi$ and $t \in \Psi_x(t) \neq \{t\}$. This is similar to case (ii). Since $\Psi_x(t) \neq \{t\}$, $x^\uparrow$ is constant in $[t - \epsilon, t)$ for all suitably small ϵ . Thus, over $(t - \epsilon, t)$, $\Phi_x^L(s)$ and $\Phi_x^R(s)$ increase to $\Psi_x^L(t)$ and $\Psi_x^R(t)$ as $s \uparrow t$. Hence, z has a left limit at t . In general, z need not be continuous at t .

Case 4(a): In this case $x^\uparrow(t) = x^\uparrow(u)$ for some $u > t$. Hence $\Phi_x^L(u) \downarrow \Phi_x^L(t)$ and $\Phi_x^R(u) \downarrow \Phi_x^R(t)$ as $u \downarrow t$ so that z is right-continuous at t . If t is not isolated in $\Phi_x(t)$, as in Case 4(b), then there exists $t_n \uparrow t$ with $x(t_n -) = x^\uparrow(t)$ or $x(t_n) = x^\uparrow(t)$, so that $x^\uparrow$ is constant in $[t - \epsilon, t]$ for all suitably small ϵ . Moreover, $\Phi_x^L(s) \uparrow \Phi_x^L(t)$ and $\Phi_x^R(s) \uparrow \Phi_x^R(t)$ as $s \uparrow t$. Hence z has a left limit $z(t-) \geq y(t-)$. Moreover, Φ_x^L and Φ_x^R are continuous at t . If $y(t-) \leq z(t-) < y(t)$, then y is right-continuous but not continuous. On the other hand, if $z(t-) \geq y(t)$, then z is continuous at t . If instead t is isolated in $\Phi_x(t)$, as in Case 4(c), then $\Phi_x^L(s)$ and $\Phi_x^R(s)$ are constant in $(t - \epsilon, t)$ for all suitably small ϵ , but $\Phi_x^R(t) = \Phi_x^R(t-) \cup \{t\}$. Hence, Φ_x^L and Φ_x^R have limits from the left at t . Thus z has a limit from the left at t , which does not depend on $y(t-)$. If $z(t-) < y(t)$, then z is discontinuous at t ; otherwise it is continuous.

Case 4(b): As in case 4(a), z has a left limit at t . If Case 4(a) does not hold, then $x^\uparrow(t) < x^\uparrow(t + \epsilon)$ for all sufficiently small ϵ . In this case, $\Phi_x(s) \rightarrow \{t\}$ as $s \downarrow t$, so that Φ_x and z have limits from the right with $z(t+) = y(t)$. However, since $\Phi_x(t) \neq \{t\}$ by assumption, Φ_x is not right-continuous. In this case z is left (right) continuous if $z(t-) \geq (\leq) y(t)$.

Case 4(c): In this case

$$t \in OK(x) \equiv Rinc(x^\uparrow) \cap Lconst(x^\uparrow) \cap Disc(x) \cap Linc(x) \cap Amax(x). \quad (3.35)$$

Note that $OK(x)$ in (3.35) differs from $Bad(x)$ in (3.28) only by having $x(t-) < x(t)$. As noted for case 4(a) and 4(b), z has left limit $z(t-)$ and right limit $z(t+) = y(t)$ at t , with $z(t) = z(t-) \vee y(t)$. However, since $x(t-) < x(t) = x^\uparrow(t)$, $t \notin \Phi_x^L(t)$, so that $z(t)$ does not depend upon $y(t-)$. Hence z is either left-continuous or right-continuous at t .

Case 4(d): In this case $t \in \text{Bad}(x)$. Since $x(t-) = x(t) = x^\uparrow(t)$, $t \in \Phi_x^L(t)$ and $z(t) \geq y(t-)$. As in Case 4(c), z has left and right limits at t with $z(t+) = y(t)$ and $z(t) = \max\{z(t-), y(t-), y(t)\}$. ■

Theorem 9.3.2 concluded that $z \in D_{\text{lim}}$ when $x, y \in D$. By the same reasoning, examining the cases in Theorem 9.3.3, we can obtain the same conclusion when $y \in D_{\text{lim}}$.

Theorem 9.3.4. (extension when $y \in D_{\text{lim}}$) *Suppose that $x \in D$ and $y \in D_{\text{lim}}$. Then $z \in D_{\text{lim}}$. At all t not in the set*

$$\text{Bad}(x, y) = [\text{Bad}_1(x) \cap \text{Disc}(y)] \cup \text{Bad}_2(y) , \quad (3.36)$$

where

$$\text{Bad}_1(x) \equiv \text{Rinc}(x^\uparrow) \cap \text{Lconst}(x^\uparrow) \cap \text{Linc}(x) \cap \text{Amax}(x) \quad (3.37)$$

and

$$\text{Bad}_2(y) \equiv \{t \in [0, T] : y(t) > y(t-), y(t+)\} , \quad (3.38)$$

z is either left-continuous or right-continuous. At $t \in \text{Bad}(x) \cap \text{Disc}(x)$, $z(t+) = y(t+)$, $z(t-)$ is independent of $y(t-)$ and $z(t) = z(t-) \vee y(t) \vee y(t+)$, so that z is left-continuous if $z(t-) \geq y(t) \vee y(t+)$, right-continuous if $y(t+) \geq y(t) \vee z(t-)$ and neither right-continuous nor left-continuous if $y(t) > z(t-) \vee y(t+)$. At $t \in \text{Bad}(x) \cap \text{Disc}(x)^c$, $z(t+) = y(t+)$, $z(t-)$ is independent of $y(t-)$ and $z(t) = z(t-) \vee y(t-) \vee y(t) \vee y(t+)$, so that z is left-continuous if $z(t-) \geq y(t-) \vee y(t) \vee y(t+)$, right-continuous if $y(t+) \geq z(t-) \vee y(t-) \vee y(t)$ and neither left-continuous nor right-continuous if $y(t-) \vee y(t) > z(t-) \vee y(t+)$.

We get extra regularity conditions if we assume that $x \in C$. Recall that z is upper semicontinuous at t if $\lim_{s \rightarrow t} z(s) \leq z(t)$; z is upper semicontinuous if it is upper semicontinuous at all t . Let D_{usc} be the subset of upper semicontinuous functions in D_{lim} .

Theorem 9.3.5. (upper-semicontinuity when $x \in C$) *Suppose that $x \in C$ and $y \in D_{\text{lim}}$, then $z \in D_{\text{usc}}$. Then $\Phi_x^L(t) = \Phi_x^R(t) = \Phi_x(t)$ for all $t > 0$ and*

$$z(t) = \sup_{s \in \Phi_x(t)} \{y(s-) \vee y(s) \vee y(s+)\} . \quad (3.39)$$

Proof. Since $x \in C$, the only relevant cases in Theorem 9.3.3 are: 1(i), 2 and 4. Formula (3.39) follows directly from formula (3.6). The upper semicontinuity follows from by considering the cases in Theorem 9.3.3. ■

Remark 9.3.3. *The need for x to be continuous.* Without assuming that $x \in C$, we need not have z be upper semicontinuous. In Case 3 of Theorem 9.3.3, we can have $z(t-) > z(t) = z(t+) = y(t+)$. ■

From the point of view of applications, the two most common cases are

$$\begin{aligned} \text{(i)} \quad & x \in C \quad \text{and} \quad y \in C \\ \text{(ii)} \quad & x \in C \quad \text{and} \quad y \in D. \end{aligned} \tag{3.40}$$

We thus summarize the situation in these two important cases.

First, with case (i) in (3.40) when both $x \in C$ and $y \in C$, we can apply Corollary 9.3.3 and Theorem 9.3.5 above to conclude that $z \in D_{l,r} \cap D_{usc}$, but Example 9.3.3 shows that we need not have $z \in D$. Indeed, we will always have $z \in D_{l,r} \cap D_{usc}$ instead. For $x \in D$, we have $x \in D_{usc}$ only if $x(t) \geq x(t-)$ for all t . So it is important to have the space $D_{l,r} \cap D_{usc}$.

Second, with Case (ii) in (3.40) when $x \in C$ but only $y \in D$, Theorem 9.3.2 shows that $z \in D_{lim}$, but Example 9.3.4 shows that we need not have $z \in D_{l,r}$ in general. However, under condition (3.29), which is implied by condition (3.30), Theorem 9.3.2 implies that we do have $z \in D_{l,r}$. Moreover Theorem 9.3.5 shows that $z \in D_{usc}$. So, in Case (ii) we should also have $z \in D_{l,r} \cap D_{usc}$, but we need to impose condition (3.30).

Because we assumed only that $y \in D_{lim}$ in Theorem 9.3.1, we can consider z playing the role of y . For example, we could start by considering $z_\epsilon(x_1, y)$ in (3.2) for some $x_1 \in D$ and obtain $z_1 = z(x, y)$ as $\epsilon \downarrow 0$. Then we could consider $z_\epsilon(x_2, z_1)$ in (3.2) for another $x_2 \in D$ and obtain $z_2 = z(x_2, z_1)$ as $\epsilon \downarrow 0$.

9.4. Extending Pointwise Convergence to M_1 Convergence

We now want to extend the pointwise convergence of z_ϵ to z as $\epsilon \downarrow 0$ in Theorem 9.3.1 to M_1 convergence. We first observe that monotone pointwise convergence of continuous functions in D does not by itself imply M_1 convergence.

Example 9.4.1. *Monotone pointwise convergence of continuous functions does not imply M_1 convergence.* To see that monotone pointwise convergence

of continuous functions does not imply M_1 convergence in $D([0, 2], \mathbb{R})$, let

$$x_{2^{-n}}(0) = x_{2^{-n}}(1 - 2^{-n}) = x_{2^{-n}}(1 - 2^{-(n+1)}) = 0$$

$$x_{2^{-n}}(1 - 3(2^{-(n+2)})) = x_{2^{-n}}(1 - 2^{-(n+1)} + 2^{-(2n+1)}) = x_{2^{-n}}(2) = 1$$

for $n \geq 1$, with $x_{2^{-n}}$ defined by linear interpolation elsewhere. Clearly $x_{2^{-n}}$ is continuous for each n . Let $x_\epsilon = x_{2^{-n}}$ for $2^{-n} \geq \epsilon > 2^{-(n+1)}$, $n \geq 1$. It is easy to see that $x_{2^{-n}}(t) \geq x_{2^{-(n+1)}}(t) \downarrow x(t)$ as $n \rightarrow \infty$ for each $t \geq 0$, so that $x_\epsilon(t) \downarrow x(t)$ as $\epsilon \downarrow 0$ for each $t \geq 0$. Moreover $x_\epsilon \rightarrow x$ in D as $\epsilon \downarrow 0$ with the M_2 topology, but not in the M_1 topology, because, for any $\delta > 0$, x_ϵ crosses the strip $(1/3, 2/3)$ for t in $[1 - \delta, 1 + \delta]$ three times for all sufficiently small ϵ , whereas x crosses it only once; see Theorem 12.5.1 (v) in the book. ■

In general (without continuity conditions) monotone pointwise convergence does not imply even M_2 convergence.

Example 9.4.2. *Monotone pointwise convergence without continuity does not imply M_2 convergence.* To see that M_2 convergence does not follow from monotone pointwise convergence in or $D_{l,r}$ when neither the limit nor the converging functions need be continuous, let $x = I_{[1,2]}$ and $x_n = 2I_{[1-n^{-1},1)} + I_{[1,2]}$, $n \geq 1$. ■

However, we can obtain a positive result when the converging functions are continuous (without relying on the special structure associated with the supremum).

Theorem 9.4.1. (M_2 convergence from monotone pointwise convergence of continuous functions) *If $x \in D_{l,r}$, $x_\epsilon \in C$ for all ϵ and $x_\epsilon(t) \downarrow x(t)$ as $\epsilon \downarrow 0$ for all $t \geq 0$, then $x_\epsilon \rightarrow x$ in $(D_{l,r}, M_2)$ as $\epsilon \downarrow 0$.*

We can combine Theorems 9.3.1 and 9.4.1 above to obtain the following corollary.

Corollary 9.4.1. (M_2 convergence of the supremum derivative) *In the setting of Theorem 9.3.1, if x and y are both continuous, then $z_\epsilon \rightarrow z$ in $(D_{l,r}, M_2)$ as $\epsilon \downarrow 0$.*

However, by exploiting the special structure of the supremum function, we will actually establish the stronger M_1 convergence under weaker conditions. To prove Theorem 9.4.1, we exploit approximations by piecewise-constant functions see Section 12.2 in the book.

Proof of Theorem 9.4.1. Since the pointwise convergence is monotone, $x_\epsilon(t) \geq x(t)$ for all t and ϵ . For any u and $\delta > 0$, let $\tilde{x}$ be a piecewise-constant function in D with $\|x - \tilde{x}\|_u < \delta$. Then $x(t) \leq \tilde{x}(t) + \delta$ for $0 \leq t \leq u$. Let $\hat{x}$ be the upper boundary (containing only vertical and horizontal pieces) of the δ neighborhood of the completed graph $\Gamma_{\tilde{x}+\delta}$ of $\tilde{x} + \delta$ for the time set $[0, t]$, using the Hausdorff metric, as depicted in Figure 9.1. Note that $\hat{x}(s) \geq x(s)$ for $0 \leq s \leq t$ and $h_t(\Gamma_x, \Gamma_{\hat{x}}) \leq 3\delta$, where h_t is the Hausdorff metric applied to the graphs with time set $[0, t]$. It thus suffices to show that $x_\epsilon(s) \leq \hat{x}(s)$ for all s , $0 \leq s \leq t$, for all sufficiently small ϵ .

Consequently, it suffices to show that $x_\epsilon(s) \vee \hat{x}(s)$ converges uniformly to $\hat{x}(s)$ for $0 \leq s \leq t$ as $\epsilon \downarrow 0$. However, $\hat{x}$ has only finitely many discontinuities. Since $x_\epsilon \vee \hat{x}$ is continuous and nonincreasing in ϵ , we can apply Dini's theorem to get uniform convergence in any compact subset of $[0, t]$ excluding arbitrarily small open neighborhoods of each of the finitely many discontinuities. To treat the discontinuities, we need to carefully treat the neighborhood to the left (right) of a jump up (down). On the other side, the limit function constrains $x_\epsilon(s) \vee \hat{x}(s)$ as $\epsilon \downarrow 0$. Now suppose that t is one of the finitely many discontinuities of $\hat{x}$. Then there is $\epsilon_0(t)$ such that $|x_\epsilon(t) - x(t)| < \delta/2$ for all $\epsilon < \epsilon_0(t)$ by the pointwise convergence. Let ϵ_0 be the minimum of the finitely many $\epsilon_0(t)$. For any $\epsilon \leq \epsilon_0$ given, the continuity of x_{ϵ_0} implies that, for each discontinuity point t , there is an $\eta(t) \equiv \eta(t, \epsilon) > 0$ such that $|x_\epsilon(t) - x_\epsilon(s)| < \delta/2$ for all s with $|s - t| < \eta(t)$. Thus, $|x_\epsilon(s) - x(t)| < \delta$ for $|s - t| < \eta(t)$. On the critical side of each discontinuity, the monotonicity implies that

$$x_{\epsilon'}(s) \leq x_\epsilon(s) \leq x_\epsilon(t) + \delta/2$$

for all $\epsilon' \leq \epsilon$. Let the open neighborhood about t be $(t - \eta(t)/4, t + \eta(t)/4)$. Outside the finite union of those open intervals, we have the uniform convergence; inside those intervals we have established that $x_\epsilon(s) \vee \hat{x}(s) < \hat{x}(s) + \delta$. Hence Γ_{x_ϵ} is contained in the 4δ -neighborhood of Γ_x for suitably small ϵ , which implies the M_2 convergence. ■

We will want to approximate $y \in D$ by $y \in D_c$. For this purpose, it is important to understand how z_ϵ and z in (3.2) and (3.6) depend upon y .

Lemma 9.4.1. (uniform Lipschitz property of z_ϵ as a function of y) For any $\epsilon > 0$, $t > 0$, $x \in D$ and $y_1, y_2 \in D$,

$$\|z_\epsilon(x, y_1) - z_\epsilon(x, y_2)\|_t \leq \|y_1 - y_2\|_t \quad (4.1)$$

and

$$\|z(x, y_1) - z(x, y_2)\|_t \leq \|y_1 - y_2\|_t. \quad (4.2)$$

Proof. Property (4.2) follows immediately from (3.6). For (4.1), note that

$$\begin{aligned}\|z_\epsilon(x, y_1) - z_\epsilon(x, y_2)\|_t &= \epsilon^{-1} \|(x + \epsilon y_1)^\uparrow - (x + \epsilon y_2)^\uparrow\|_t \\ &\leq \epsilon^{-1} \|(x + \epsilon y_1) - (x + \epsilon y_2)\|_t \\ &= \|y_1 - y_2\|_t \cdot \blacksquare\end{aligned}$$

We also employ the following elementary, but useful, lemma.

Lemma 9.4.2. ($z \in D_c$ when $y \in D_c$) *Suppose that $x, y \in D$. If, in addition, $y \in D_c$ and*

$$Disc(y) \cap Bad(x) = \emptyset \quad (4.3)$$

for $Bad(x)$ in (3.28), then $z \in D_c$. If y has k discontinuity points in $(0, t)$, then z has at most k discontinuity points in $[0, t]$.

Proof. We use Theorem 9.3.2 to show that $z \in D$. Since $y \in D_c$, for any given interval $[0, t]$, there are time points $t_0 = 0 < t_1 < \dots < t_k = t$ such that y is constant on $[t_{j-1}, t_j]$ and $[t_{k-1}, t]$ for $1 \leq j \leq k$. Note that $z(t) = y(0)$ for $t \in [0, t_1)$. From (3.6), it is obvious that z can only assume one of the k values $y(t_{j-1})$, $1 \leq j \leq k$. The function z may change to $y(t_{j-1})$ in the interval $[t_{j-1}, t_j]$, but it can only do so once. Transitions from $z(t_{j-1}-) < y(t_{j-2})$ to $y(t_{j-2}) < y(t_{j-1})$ to $y(t_{j-1})$ at t_{j-1} are ruled out by condition (4.3). $\blacksquare$

Theorem 9.4.2. (M_1 convergence of the supremum derivative) *Suppose that $x, y \in D$ and (4.3) holds for $Bad(x)$ in (3.28). Then*

$$z_\epsilon \rightarrow z \quad \text{in } (D_{lr}, M_1) \quad \text{as } \epsilon \downarrow 0$$

for z_ϵ in (3.2) and z in (3.6).

Proof. Lemmas 9.4.1 and 9.4.2 imply that it suffices to consider $y \in D_c$ in order to establish the M_1 convergence. By Theorem 9.3.2 and Example 9.3.2, the discontinuity condition (4.3) is necessary and sufficient to have $z \in D$. Under condition (4.3), it is possible to choose the piecewise-constant approximation to y so that it too satisfies (4.3). So, henceforth, assume that $y \in D_c$ and satisfies (4.3). By Lemma 9.4.2, $z \in D_c$ as well. Now, by applying mathematical induction over the successive discontinuities of z , it is not difficult to show that, for all sufficiently small $\epsilon > 0$, $z_\epsilon(t) = z(t)$ for all t outside a union of open neighborhoods of the discontinuities of z . (We strongly exploit D_c at this step.) For given discontinuities of y and z , by

making ϵ suitably small, these neighborhoods can be chosen to be disjoint with the property that z_ϵ is monotone on each interval. The monotonicity together with the pointwise convergence established in Theorem 9.3.1 implies the local characterization of M_1 convergence in Theorem 12.5.1 in the book. ■

Example 9.4.3. *The need for M_1 convergence.* It is possible to have $z_\epsilon = z$ at a discontinuity point of z : For $x(t) = 0$, $t \geq 0$, $z_\epsilon(t) = z(t) = y^\uparrow(t)$ for all $t \geq 0$. Then z_ϵ and z have the discontinuities of $y^\uparrow$. A typical case requiring the M_1 convergence is $y = I_{[1,\infty)}$ and $x(t) = -tI_{[0,1)}(t) + (t-2)I_{[1,\infty)}(t)$. Then

$$z_\epsilon(t) = \epsilon^{-1}(2-t+\epsilon)I_{[2-\epsilon,2)}(t) + I_{[2,\infty)}(t) \rightarrow z(t) = I_{[2,\infty)}(t) \quad \text{in } (D, M_1) .$$

Finally, we can combine Theorems 9.2.3, 9.4.2 and the triangle inequality (2.1) to obtain a preservation-of-convergence result for the supremum function.

Theorem 9.4.3. (convergence preservation for the supremum map with nonlinear centering) *For $\epsilon > 0$, let $x_\epsilon, y \in D$ and let x be a Lipschitz function in C . If*

$$d_{M_1}(x_\epsilon - x, \epsilon y) = o(\epsilon) \quad \text{as } \epsilon \downarrow 0 , \quad (4.4)$$

for which a sufficient condition is

$$\|\epsilon^{-1}(x_\epsilon - x) - y\|_t \rightarrow 0 \quad \text{as } \epsilon \downarrow 0 \quad \text{for all } t > 0 , \quad (4.5)$$

and if (4.3) holds for $\text{Bad}(x)$ in (3.28), then

$$\epsilon^{-1}(x_\epsilon^\uparrow - x^\uparrow) \rightarrow z \quad \text{in } (D_{l,r}, M_1) \quad \text{as } \epsilon \downarrow 0 \quad (4.6)$$

for z in (3.6).

Corollary 9.4.2. (convergence preservation starting with the standard initial limit (4.5)) *For $\epsilon > 0$, let $x_\epsilon \in D$ and $x, y \in C$ with x being Lipschitz. If (4.5) holds, then (4.6) holds for z in (3.39) and $z \in D_{usc} \cap D_{l,r}$.*

9.5. Derivative of the Reflection Map

Now we consider the reflection map $\phi : D \rightarrow D$ defined by

$$\phi(x) \equiv x + (-x \vee 0)^\uparrow ; \quad (5.1)$$

see Section 13.4 in the book.

Results for the reflection map ϕ in (5.1) above follow from the results for the supremum map in Sections 9.3 and 9.4 above, because

$$\phi_\epsilon(x, y) \equiv \epsilon^{-1}[\phi(x + \epsilon y) - \phi(x)] = y + m_\epsilon(-x, -y) , \quad (5.2)$$

where

$$m_\epsilon(x, y) = \epsilon^{-1}[(x + \epsilon y)^\uparrow \vee 0 - (x^\uparrow \vee 0)] . \quad (5.3)$$

Note that $m_\epsilon(x, y)$ in (5.3) differs from $z_\epsilon(x, y)$ in (3.2) only by the extra maximum with respect to 0. In most applications, we will have $x(0) = y(0) = 0$, in which case the extra maximum $\vee 0$ is superfluous; then $m_\epsilon(x, y) = z_\epsilon(x, y)$. Thus, in this common case we can immediately apply the results in Section 9.3 to obtain corresponding results for the reflection map.

Theorem 9.5.1. (derivative of the reflection map in the common case) *Suppose that $x \in D$, $y \in D$ and $x(0) = y(0) = 0$. Then, for each $t > 0$,*

$$\lim_{\epsilon \downarrow 0} \phi_\epsilon(x, y)(t) = \dot{\phi}(t) , \quad (5.4)$$

where

$$\begin{aligned} \dot{\phi}(t) &\equiv \dot{\phi}(x, y)(t) \\ &\equiv y(t) - \left(\inf_{s \in \Phi_{-x}^L(t)} \{y(s-)\} \wedge \inf_{s \in \Phi_{-x}^R(t)} \{y(s)\} \right) \end{aligned} \quad (5.5)$$

and $\dot{\phi} \in D_{lim}$. If, in addition,

$$Disc(y) \cap Bad(-x) = \emptyset , \quad (5.6)$$

then $\dot{\phi} \in D_{l,r}$ and

$$\phi_\epsilon(x, y) \rightarrow \dot{\phi}(x, y) \quad \text{in } (D_{l,r}, M_1) \quad \text{as } \epsilon \downarrow 0 . \quad (5.7)$$

If, in addition, $x \in C$, then $\dot{\phi} \in D_{usc}$. If, in addition, x is Lipschitz and $y \in C$, then there is convergence preservation: If

$$\|\epsilon^{-1}(x_\epsilon - x) - y\|_t \rightarrow 0 \quad \text{as } \epsilon \downarrow 0 \quad \text{for all } t \quad (5.8)$$

then

$$\epsilon^{-1}(\phi(x_\epsilon) - \phi(x)) \rightarrow \dot{\phi}(x, y) \quad \text{in } (D_{l,r}, M_1) \quad \text{as } \epsilon \downarrow 0 . \quad (5.9)$$

for

$$\dot{\phi}(t) = y(t) - \inf_{s \in \Phi_{-x}(t)} \{y(s)\} . \quad (5.10)$$

where

$$\Phi_{-x}(t) = \{s : 0 \leq s \leq t, x(s) = x^\downarrow(t)\} , \quad t \geq 0 . \quad (5.11)$$

Proof. The pointwise limit in (5.4) follows from Theorem 9.3.1, noting that $-(-y)^\uparrow = y^\downarrow$. The fact that $\phi \in D_{lim}$ follows from Theorem 9.3.2. The stronger conclusion that $\phi \in D_{l,r}$ under condition (5.6) also follows from Theorem 9.3.2, exploiting condition (3.30). The M_1 convergence in (5.7) follows from Theorem 9.4.2. Finally, the convergence preservation ((5.8) implies (5.9)) follows from Corollary 9.4.3. ■

We now return to the general case. For that purpose, let

$$t_l \equiv t_l(x) \equiv \inf\{t > 0 : x^\uparrow(t) = 0\} \quad (5.12)$$

and

$$t_u = t_u(x) \equiv \sup\{t > 0 : x^\uparrow(t) = 0\}, \quad (5.13)$$

with $t_l = t_u = \infty$ if $x^\uparrow(t) < 0$ for all t . In many applications we will have $x(0) = 0$; then $t_l = 0$ and $t_u = \infty$. It is easy to see that for any t , $0 \leq t < t_l$, $m_\epsilon(x, y)(t) = 0$ for all sufficiently small positive ϵ . Similarly, for any t , $t_u < t < \infty$, $m_\epsilon(x, y)(t) = z_\epsilon(x, y)(t)$ for all sufficiently small positive ϵ . We need to examine the interval $(t_l - \epsilon, t_u + \epsilon)$ more carefully. To do so, we exploit the following analog of Lemma 9.4.1, which is proved in the same way.

Lemma 9.5.1. (uniform Lipschitz property for m_ϵ as a function of y) *For any $\epsilon > 0$, $t > 0$, $x \in D$ and $y_1, y_2 \in D$,*

$$\|m_\epsilon(x, y_1) - m_\epsilon(x, y_2)\|_t \leq \|y_1 - y_2\|_t.$$

Our analog of Theorems 9.3.1, 9.3.2, 9.3.5 and 9.4.2 for m_ϵ is the following.

Theorem 9.5.2. (the derivative in the general case) *Suppose that $x, y \in D$. For each $t \geq 0$, $m_\epsilon(x, y)(t)$ is decreasing in ϵ and*

$$\lim_{\epsilon \downarrow 0} m_\epsilon(x, y)(t) = m(x, y)(t) \equiv \begin{cases} 0, & t < t_l \\ y(t-) \vee y(t) \vee 0, & t = t_l \\ z(t) \vee 0, & t_l < t < t_u \\ z(t-) \vee 0 \vee y(t), & t = t_u \\ z(t), & t > t_u \end{cases} \quad (5.14)$$

for m_ϵ in (5.3), t_l in (5.12), t_u in (5.13) and $z(t)$ in (3.6). The limit $m(x, y)$ in (5.14) has limits from the left and right at all t . If $x \in C$, then z is given by (3.39) and z and m are upper semicontinuous. At all t not in the set

$$B(x) \equiv \{t_l\} \cup (Bad(x) \cap (t_l, \infty)) \quad (5.15)$$

for $\text{Bad}(x)$ in (3.28), m is either left-continuous or right-continuous. At $t = t_l$, m is left-continuous if $y(t-) \vee y(t) \leq 0$, m is right-continuous if $y(t) \geq y(t-) \vee 0$, and neither left-continuous nor right-continuous if $y(t-) > y(t) \vee 0$. If

$$(i) \ y(t-) \leq z(t-) \vee y(t) \vee 0 \quad \text{for } t \in B(x) \cap [t_l, t_u] \quad (5.16)$$

and

$$(ii) \ y(t-) \leq z(t-) \vee y(t) \quad \text{for } t \in B(x) \cap (t_u, \infty) , \quad (5.17)$$

for which a sufficient condition is

$$\text{Disc}(y) \cap B(x) = \emptyset , \quad (5.18)$$

then m is either left-continuous or right-continuous at all t , so that $m \in D_{l,r}$. Then

$$m_\epsilon(x, y) \rightarrow m(x, y) \quad \text{in } (D_{l,r}, M_1) \quad \text{as } \epsilon \downarrow 0 .$$

Proof. First, for any $\delta > 0$ and $T > 0$, $m_\epsilon(x, y)(t) = 0$ in $[0, (0 \vee (t_l - \delta)) \wedge T]$ and $m_\epsilon(x, y)(t) = z_\epsilon(x, y)(t)$ in $[(t_u + \delta) \wedge T, T]$ for all sufficiently small positive ϵ . We apply Theorems 9.3.1, 9.3.2 and 9.4.2 to treat the subinterval $[(t_u + \delta) \wedge T, T]$. Hence it suffices to focus on the subinterval $(t_l - \delta, t_u + \delta)$. By Lemmas 9.4.1 and 9.5.1, it suffices to assume that $y \in D_c$. The argument then is as for Theorems 9.3.1, 9.3.2, 9.3.5 and 9.4.2. ■

Corollary 9.5.1. (convergence) *If $x, y \in D$, then*

$$\phi_\epsilon(x, y)(t) \downarrow y(t) + m(-x, -y)(t) \quad \text{as } \epsilon \downarrow 0$$

for ϕ_ϵ in (5.2), each $t \geq 0$ and m in (5.14). If in addition (5.18) holds, then

$$\phi_\epsilon(x, y) \rightarrow y + m(-x, -y) \quad \text{in } (D_{l,r}, M_1) \quad \text{as } \epsilon \downarrow 0 .$$

Finally, paralleling Theorem 9.4.3 for the supremum function, we can combine Theorems 9.2.3, 9.5.2 and the triangle inequality in (2.1) to obtain a preservation-of-convergence result for the reflection map.

Theorem 9.5.3. (M_1 convergence for the reflection derivative) *For $\epsilon > 0$, let $x_\epsilon, y \in D$ and let x be a Lipschitz function in C . If condition (4.4) holds, for which a sufficient condition is (4.5), and if (5.18) holds, then*

$$\epsilon^{-1}(\phi(x_\epsilon) - \phi(x)) \rightarrow y + m(-x, -y) \quad \text{in } (D_{l,r}, M_1) \quad \text{as } \epsilon \downarrow 0 \quad (5.19)$$

for m in (5.14).

Corollary 9.5.2. *For $\epsilon > 0$, let $x_\epsilon \in D$ and $x, y \in C$ with x being Lipschitz. If (4.5) holds, then (5.19) holds for m in (5.14), where $m \in D_{usc} \cap D_{l,r}$.*

9.6. Heavy-Traffic Limits for Nonstationary Queues

In this section we apply the convergence-preservation results in the last section to establish heavy-traffic limits for nonstationary queues. We assume that the queue-length process can be represented directly as the reflection map applied to a net-input process, which is the difference of two nondecreasing processes admitting nonstationary rates.

As background, note that the queue-length process $\{Q(t) : t \geq 0\}$ in the M/M/1 queue starting empty with arrival rate λ and service rate μ has such a representation. In particular, for the M/M/1 queue,

$$Q(t) = \phi(X)(t), \quad t \geq 0, \quad (6.1)$$

where X is the net-input process, satisfying

$$X(t) = X^+(\Lambda^+(t)) - X^-(\Lambda^-(t)), \quad (6.2)$$

with X^+ and X^- being rate-1 Poisson processes and

$$\Lambda^+(t) = \lambda t \quad \text{and} \quad \Lambda^-(t) = \mu t, \quad t \geq 0. \quad (6.3)$$

Then $X^+ \circ \Lambda^+$ is a rate- λ Poisson process.

Similarly, for the $M_t/M_t/1$ queue with (integrable) time-dependent arrival-rate function $\lambda(t)$ and service-rate function $\mu(t)$, (6.1) and (6.2) remain valid with Λ^+ and Λ^- redefined as

$$\Lambda^+(t) = \int_0^t \lambda(s) ds \quad \text{and} \quad \Lambda^-(t) = \int_0^t \mu(s) ds. \quad (6.4)$$

It is easy to see that there are many generalizations. First, we obtain the queue-length process in an MMPP/MMPP/1 queue with independent Markov modulated Poisson process (MPPP) arrival and service processes if Λ^+ and Λ^- are independent stationary versions of finite-state continuous-time Markov chains. (We then assume that X^+ , X^- , Λ^+ and Λ^- are mutually independent. We obtain the queue-length process in a more general MMPP_t/MMPP_t/1 queue with independent time-dependent MMPP arrival and service processes if Λ^+ and Λ^- are independent time-dependent finite-state CTMCs, governed by time-dependent transition functions.

We construct associated fluid queue models by letting X^+ and X^- be other Lévy processes instead of Poisson processes. Without loss of generality, these again can be rate-1 processes. For nodes in a communication network with fixed bandwidth, it is natural to let $X^-(t) = t$, $t \geq 0$, but generalizations are possible.

We now establish limits for a sequence of models indexed by n . For each n , we have the four-tuple of stochastic processes $(X_n^+, X_n^-, \Lambda_n^+, \Lambda_n^-)$ with sample paths in D^4 . We then form the associated scaled stochastic processes by letting

$$\begin{aligned} \mathbf{X}_n^+(t) &\equiv c_n^{-1}[X_n^+(nt) - nx^+(t)] \\ \mathbf{X}_n^-(t) &\equiv c_n^{-1}[X_n^-(nt) - nx^-(t)] \\ \Lambda_n^+(t) &\equiv c_n^{-1}[\Lambda_n^+(t) - ny^+(t)] \\ \Lambda_n^-(t) &\equiv c_n^{-1}[\Lambda_n^-(t) - ny^-(t)] \\ \mathbf{X}_n(t) &\equiv c_n^{-1}[X_n^+(\Lambda_n^+(t)) - X_n^-(\Lambda_n^-(t)) - nx^+(y^+(t)) - x^-(y^-(t))] \\ \hat{\mathbf{X}}_n(t) &\equiv n^{-1}[X_n^+(\Lambda_n^+(t)) - X_n^-(\Lambda_n^-(t))], \quad t \geq 0, \end{aligned} \quad (6.5)$$

We think of the centering terms x^+ , x^- , y^+ and y^- as deterministic functions, but that is not necessary.

The following limit for the net-input process is a direct consequence of Theorem 13.3.2 in the book.

Theorem 9.6.1. (FLLN and FCLT for the net-input process) *Suppose that*

$$(\mathbf{X}_n^+, \mathbf{X}_n^-, \Lambda_n^+, \Lambda_n^-) \Rightarrow (\mathbf{U}^+, \mathbf{U}^-, \mathbf{V}^+, \mathbf{V}^-) \quad \text{in} \quad (D^4, WM_1) \quad (6.6)$$

for the processes in (6.5), where x^+ and x^- have continuous derivatives $\dot{x}^+$ and $\dot{x}^-$, y^+ and y^- are continuous nonnegative and strictly increasing, $c_n \rightarrow \infty$, $n/c_n \rightarrow \infty$ and

$$\begin{aligned} \text{Disc}(\mathbf{U}^+ \circ y^+) \cap \text{Disc}(\mathbf{V}^+) &= \phi \\ \text{Disc}(\mathbf{U}^- \circ y^-) \cap \text{Disc}(\mathbf{V}^-) &= \phi \\ \text{Disc}(\mathbf{U}^+ \circ y^+ + (\dot{x}^+ \circ y^+) \mathbf{V}^+) \cap \\ \text{Disc}(\mathbf{U}^- \circ y^- + (\dot{x}^- \circ y^-) \mathbf{V}^-) &= \phi. \end{aligned} \quad (6.7)$$

Then

$$\hat{\mathbf{X}}_n \Rightarrow x \quad \text{in} \quad (D, M_1) \quad (6.8)$$

and

$$\mathbf{X}_n \Rightarrow \mathbf{X} \quad \text{in} \quad (D, M_1), \quad (6.9)$$

for $\hat{\mathbf{X}}_n$ and $\mathbf{X}_n$ in (6.5), where

$$x \equiv x^+ \circ y^+ - x^- \circ y^- \quad (6.10)$$

and

$$\mathbf{X} \equiv \mathbf{U}^+ \circ y^+ + (\dot{x}^+ \circ y^+) \mathbf{V}^+ - \mathbf{U}^- \circ y^- - (\dot{x}^- \circ y^-) \mathbf{V}^-. \quad (6.11)$$

Proof. As usual, start by applying the Skorohod representation theorem to replace the convergence in distribution in (6.6) by convergence w.p.1 for special versions, without introducing new notation for the special versions. Then apply Theorem 13.3.2 in the book, after rewriting $\mathbf{X}_n^+$ as

$$\mathbf{X}_n^+(t) \equiv (n/c_n)[n^{-1}X_n^+(nt) - x^+(t)], \quad t \geq 0, \quad (6.12)$$

and similarly for the other functions. That yields

$$\begin{aligned} & c_n^{-1}(X_n^+ \circ \Lambda_n^+ - nx^+ \circ y^+, X_n^- \circ \Lambda_n^- - nx^- \circ y^-) \\ & \Rightarrow (U^+ \circ y^+ + (\dot{x}^+ \circ y^+)V^+, U^- \circ y^- + (\dot{x}^- \circ y^-)V^-) \end{aligned} \quad (6.13)$$

in (D^2, WM_1) . Multiply by c_n/n in (6.13) to get

$$n^{-1}(X_n^+ \circ \Lambda_n^+, X_n^- \circ \Lambda_n^-) \Rightarrow (x^+ \circ y^+, x^- \circ y^-) \quad \text{in } (D^2, WM_1) \quad (6.14)$$

Finally, given the last condition in (6.7), we can apply addition to go from (6.13) and (6.14) to (6.9) and (6.8). ■

We now apply Theorem 9.5.1 to obtain a corresponding result for the queue-length processes. Let

$$\mathbf{Q}_n(t) \equiv c_n^{-1}(Q_n(nt) - nq(t)), \quad t \geq 0. \quad (6.15)$$

and

$$\hat{\mathbf{Q}}_n(t) \equiv n^{-1}Q_n(nt), \quad t \geq 0. \quad (6.16)$$

Theorem 9.6.2. (FLLN and FCLT for the queue-length process) *If, in addition to the assumptions of Theorem 9.6.1, y^+ and y^- are Lipschitz continuous, $x(0) = 0$, $P(\mathbf{X}(0) = 0) = 1$ and*

$$P((\mathbf{U}^+, \mathbf{U}^-, \mathbf{V}^+, \mathbf{V}^-) \in C^4) = 1, \quad (6.17)$$

then

$$\hat{\mathbf{Q}}_n \Rightarrow q \quad \text{in } (D, M_1) \quad (6.18)$$

and

$$\mathbf{Q}_n \Rightarrow \mathbf{Q} \quad \text{in } (D_{l,r}, M_1) \quad (6.19)$$

for $\hat{\mathbf{Q}}_n$ in (6.16) and $\mathbf{Q}_n$ in (6.15), where

$$q = \phi(x) \quad (6.20)$$

for x in (6.10) and

$$\mathbf{Q} = \mathbf{X} + z(-x, -\mathbf{X}) \quad (6.21)$$

for x in (6.10), $\mathbf{X}$ in (6.11) and z in (3.16). The limit process $\mathbf{Q}$ then has upper semicontinuous sample paths.

Example 9.6.1. *The $M_t/M_t/1$ queue.* Now let us examine the special case of the $M_t/M_t/1$ queue in more detail. For the $M_t/M_t/1$ queue, $c_n = \sqrt{n}$, $x^+ = x^- = e$ and U^+, U^- are independent Brownian motions. It is natural to have

$$\Lambda_n^+(t) = \int_0^t \lambda_n^+(s) ds \quad \text{and} \quad y^\pm(t) = \int_0^t \lambda^\pm(s) ds \quad (6.22)$$

where $\lambda_n^\pm$ and $\lambda^\pm$ are deterministic functions. We can then have

$$n^{-1/2}(\lambda_n^\pm(t) - n\lambda^\pm(t)) \rightarrow \gamma^\pm(t) \quad \text{as} \quad n \rightarrow \infty \quad (6.23)$$

uniformly in $[0, T]$, where γ^+ and γ^- are deterministic, which implies that

$$\Lambda_n^\pm(t) \rightarrow \int_0^t \gamma^\pm(s) ds \equiv \mathbf{V}^\pm. \quad (6.24)$$

Thus the assumptions of Theorems 9.6.1 and 9.6.2 are satisfied and

$$x(t) = \int_0^t [\lambda^+(s) - \lambda^-(s)] ds, \quad t \geq 0, \quad (6.25)$$

while

$$\begin{aligned} \mathbf{X}(t) = & \mathbf{U}^+ \left(\int_0^t \lambda^+(s) ds \right) \\ & - \mathbf{U}^- \left(\int_0^t \lambda^-(s) ds \right) + \int_0^t [\gamma^+(s) - \gamma^-(s)] ds \end{aligned} \quad (6.26)$$

where $\mathbf{U}^+$ and $\mathbf{U}^-$ are independent standard Brownian motions and the rest involves continuous deterministic functions. It is easy to see that X is equal in distribution (on D) to

$$U \left(\int_0^t [\lambda^+(s) + \lambda^-(s)] ds \right) + \int_0^t [\gamma^+(s) - \gamma^-(s)] ds, \quad t \geq 0, \quad (6.27)$$

where U is a standard Brownian motion.

The FWLLN limits x and q can be regarded as the net-input and buffer-content processes, respectively, in a fluid-queue model with time-dependent deterministic input rate $\lambda^+(t)$ and time-dependent deterministic potential output rate $\lambda^-(t)$. Then

$$-(-x)^\downarrow = - \min_{0 \leq s \leq t} \left\{ \int_0^s [\lambda^-(r) - \lambda^+(r)] dr \right\} \quad (6.28)$$

represents the cumulative potential output that is lost (i.e., does not occur during the interval $[0, t]$ because of insufficient input. Then

$$\Phi_{-x}(t) = \{s : 0 \leq s \leq t, q(s) = 0, -(-x)^\downarrow(s) = -(-x)^\downarrow(t)\} \quad (6.29)$$

i.e., $\Phi_{-x}(t)$ is the set of times s at which the buffer is empty ($q(s) = 0$) and there is no potential output loss over $[s, t]$.

An important special case is when λ_n^+ and λ_n^- in (6.22) are independent of n . Then $\gamma^+(t) = \gamma^-(t) = 0$ for all $t \geq 0$ and the deterministic function $\int_0^t [\gamma^+(s) - \gamma^-(s)] ds$ in (6.27) is identically 0. Then the limit for the queue-length process has one of three forms over subintervals: time-scaled standard Brownian motion (BM), time-scaled canonical reflected Brownian motion (RBM) and the zero function. There can be discontinuities in the sample path when the set function $\Phi_{-x}(t)$ is discontinuous in t . We display possible sample paths of (λ^+, λ^-) , $(-x, (-x)^\uparrow)$, $\Phi_{-x}(t)$, q and Q when λ^- is the constant function in Figure 9.2 below. We identify nine intervals associated with nine time points $t_0 \equiv 0 < t_1 < \dots < t_8$.

In this example, the fluid rates start out ordered by $\lambda^+(t) < \lambda^-(t)$. Thus $-x(t) \equiv \int_0^t [\lambda^-(s) - \lambda^+(s)] ds$ is initially increasing, which implies that $\Phi_{-x}(t) = \{t\}$. Thus $Q(t) = q(t) = 0$ for these t . At time t_1 , the ordering switches to $\lambda^+(t) > \lambda^-(t)$. Thus after t_1 , $-x$ is decreasing, so that $\Phi_{-x}(t) = \{t_1\}$. At time t_2 , the ordering switches back to $\lambda^+(t) < \lambda^-(t)$, but $-x(t)$ does not reach $(-x)^\uparrow(t) = (-x)(t_1)$ and $q(t)$ does not return to 0 until $t = t_3$. In the interval (t_1, t_3) , q is positive and Q is time-scaled BM.

At time t_3 , there is a discontinuity in the set-valued function Φ_{-x} and a corresponding jump in the stochastic process Q . In the interval (t_3, t_4) , $-x$ is still increasing and $\Phi_{-x}(t) = \{t\}$, so that $q(t) = Q(t) = 0$, just as in $[0, t_1)$. In the interval (t_4, t_5) , $\lambda^+(t) = \lambda^-$, so that $-x$ is constant and $\Phi_{-x}(t) = [t_4, t]$, $t_4 \leq t \leq t_5$. In the interval (t_4, t_5) , Q evolves as RBM. At t_5 , λ^+ increases, so that $-x$ decreases and $\Phi_{-x}(t) = \Phi_{-x}(t_5) = [t_4, t_5]$ for $t_5 \leq t < t_7$. At t_6 , λ^+ starts to decrease again and at t_7 $q(t) = 0$ for the first time. Hence, Q evolves as BM in the interval (t_5, t_7) .

At t_7 , there is a second discontinuity in Φ_{-x} and a corresponding jump in Q . In the subsequent interval $[t_7, t_8]$, $\lambda^+(t) = \lambda^-$, so that $-x$ remains constant. Then $\Phi_{-x}(t) = [t_4, t_5] \cup [t_7, t]$ for $t_7 \leq t < t_8$. During the interval $[t_7, t_8]$, $q(t) = 0$ and Q evolves as RBM. At t_8 , λ^+ starts to decrease and thereafter remains below λ^- . Hence, Φ_{-x} has another discontinuity at t_8 . After t_8 , $\Phi_{-x}(t) = \{t\}$ and $q(t) = Q(t) = 0$.

We conclude this section by relating the three possible kinds of heavy-traffic limits for the case of the $M_t/M_t/1$ queue with fixed arrival and service

rate functions $\lambda^+(t)$ and $\mu^-(t)$ to the values of a *time-dependent traffic intensity*, defined by

$$\rho^*(t) \equiv \sup_{0 \leq s \leq t} \left\{ \int_0^t \lambda^+(r) dr / \int_s^t \lambda^-(r) dr \right\}, \quad t \geq 0. \quad (6.30)$$

Notice that the buffer-content deterministic fluid limit q satisfies

$$\begin{aligned} q(t) &= x(t) - \inf_{0 \leq s \leq t} x(s) \\ &= \sup_{0 \leq s \leq t} \{x(t) - x(s)\} \\ &= \sup_{0 \leq s \leq t} \left\{ \int_s^t [\lambda^+(r) - \lambda^-(r)] dr \right\}, \end{aligned} \quad (6.31)$$

so that $q(t) > 0$ if and only if $\rho^*(t) > 1$.

Moreover, we can have $q(t) = 0$ but $P(Q(t) = 0) = 0$ for all t in an interval (a, b) if and only if $\rho^*(t) = 1$ in (a, b) . First, we must have $\rho^* \leq 1$ since $q(t) = 0$. However, in this region we must also have

$$\int_s^t [\lambda^+(r) - \lambda^-(r)] dr = 0 \quad (6.32)$$

for some s suitably chose to t . For that s ,

$$\int_s^t \lambda^+(r) dr / \int_s^t \lambda^-(r) dr = 1 \quad (6.33)$$

which implies that $\rho^*(t) \geq 1$. Since both $\rho^*(t) \leq 1$ and $\rho^*(t) \geq 1$, we must have $\rho^*(t) = 1$.

We thus say that the queue is *overloaded*, *critically loaded* or *underloaded* in an open interval (a, b) if $\rho^*(t) > 1$, $\rho^*(t) = 1$ or $\rho^*(t) < 1$ throughout the interval (a, b) . In Figure 9.2 above, in the intervals $(0, t_1)$, (t_1, t_3) , (t_3, t_4) , (t_4, t_5) , (t_5, t_7) , (t_7, t_8) and (t_8, T) , we have successively $\rho^*(t) < 1$, > 1 , < 1 , $= 1$, > 1 , $= 1$ and < 1 .

9.7. Derivative of the Inverse Map

In this section we obtain convergence-preservation results for the inverse map

$$x^{-1}(t) \equiv \inf\{s \geq 0 : x(s) > t\}, \quad t \geq 0, \quad (7.1)$$

defined on the subset D_u of functions unbounded above in $D \equiv D([0, \infty), \mathbb{R})$, as in Section 13.6 of the book. As in previous sections here, we approach convergence preservation through a derivative representation.

To determine the derivative of the inverse map, we introduce yet another topology on D . Recall that we introduced the M'_1 topology on $D([0, t], \mathbb{R})$ by appending a segment to the graphs, i.e., by letting

$$\Gamma'_x = \Gamma_x \cup \{(\alpha x(0), 0) : 0 \leq \alpha \leq 1\}, \quad (7.2)$$

where Γ_x is the graph of x , i.e.,

$$\begin{aligned} \Gamma_x &\equiv \{(z, s) \in \mathbb{R} \times [0, t] : \\ &z = \alpha x(s-) + (1 - \alpha)x(s) \text{ for some } \alpha, 0 \leq \alpha \leq 1\}. \end{aligned} \quad (7.3)$$

We now construct a similar M''_1 topology on $D([0, t], \mathbb{R})$ by also appending the vertical line at t to the graph, i.e., by setting

$$\Gamma''_x = \Gamma'_x \cup (\mathbb{R} \times \{t\}) \quad (7.4)$$

for Γ'_x in (7.2). Note that the function value at the right endpoint t plays no role in the M''_1 topology.

As done before for the graph Γ_x in (7.3), we define a lexicographic *order relation* on the graph Γ''_x by saying that $(z_1, s_1) \leq (z_2, s_2)$ if either (i) $s_1 < s_2$ or (ii) $s_1 = s_2$ and $|x(s_1-) - z_1| \leq |x(s_1-) - z_2|$. The definition makes the relation $\leq$ a total order on the graph Γ''_x . A parametric representation of the graph Γ''_x or the function x is a continuous nondecreasing function (u, r) mapping $[0, 1]$ into the graph Γ''_x such that $r(0) = 0$, $u(0) = 0$ and $r(1) = t$. We allow the parametric representation of Γ''_x to cover only part of the vertical line at t . If $r(s) < t$ for all $s < 1$, then the parametric representation (u, r) covers only the single point $(x(t-), t)$. If $r(s) = t$ for $a \leq s \leq 1$, then (u, r) covers a compact subinterval of either $\{(z, t) : z \geq x(t-)\}$ or $\{(z, t) : z \leq x(t-)\}$. (Since (u, r) maps $[0, 1]$ into Γ''_x , we must have $(u(1), r(1)) \in \Gamma''_x$, which implies that $|u(1)| < \infty$.) Let $\Pi''(x)$ be the set of all parametric representations of Γ''_x .

A metric d''_t inducing the M''_1 topology on $D([0, t], \mathbb{R})$ is defined by letting

$$d''_t(x_1, x_2) = \inf_{\substack{(u_i, r_i) \in \Pi''(x_i) \\ i=1,2}} \{\|u_1 - u_2\|_1 \vee \|r_1 - r_2\|_1\}. \quad (7.5)$$

We have the following lemma linking the M'_1 and M''_1 topologies with bounded function domains.

Lemma 9.7.1. *Let $x, x_n \in D([0, \infty), \mathbb{R})$. If $x_n \rightarrow x$ as $n \rightarrow \infty$ for the restrictions in $D([0, t_2], \mathbb{R}, M_1'')$ for $0 < t_2 < \infty$, then $x_n \rightarrow x$ as $n \rightarrow \infty$ for the restrictions in $D([0, t_1], \mathbb{R}, M_1')$ for each $t_1 \in \text{Disc}(x)^c$ with $0 < t_1 < t_2$.*

As before, we say that $x_n \rightarrow x$ in $D([0, \infty), \mathbb{R})$ with any of the topologies M_1 , M_1' or M_1'' if $x_n \rightarrow x$ for the restrictions in $D([0, t], \mathbb{R})$ ($D([0, t], \mathbb{R})$ for M_2'') with the same topology for all t in a sequence $\{t_k\}$ with $t_k \rightarrow \infty$ as $k \rightarrow \infty$. (The boundary points t_k can be taken from $\text{Disc}(x)^c$.) We obtain the following result from Lemma 9.7.1.

Lemma 9.7.2. *The M_1' and M_1'' topologies coincide on $D([0, \infty), \mathbb{R})$.*

We can combine Lemma 9.7.2 here and Theorem 13.6.3 in the book to obtain the following connection between M_1'' and M_1 .

Lemma 9.7.3. *If*

$$x_n \rightarrow x \quad \text{in} \quad D([0, \infty), \mathbb{R}, M_1''),$$

where $x(0) = 0$, then

$$x_n \rightarrow x \quad \text{in} \quad D([0, \infty), \mathbb{R}, M_1).$$

A metric d'' inducing the M_1'' topology on $D([0, \infty), \mathbb{R})$ is defined by letting

$$d''(x_1, x_2) = \int_0^\infty e^{-t} [1 \wedge d_t''(x_1, x_2)] dt, \quad (7.6)$$

where $d_t''(x_1, x_2)$ is understood to be the d_t'' metric applied to the restrictions of x_1 and x_2 to $[0, t]$. There is convergence $d''(x_n, x) \rightarrow 0$ if and only if there exist parametric representations (u, r) of x and (u_n, r_n) of x_n , $n \geq 1$ with domains $[0, \infty)$, such that $\|u_n - u\|_t \vee \|r_n - r\|_t \rightarrow 0$ as $n \rightarrow \infty$ for each t .

To apply the approach in Section 9.2, we need the inverse map to be Lipschitz. The Lipschitz property is valid on an appropriate subset of D with an appropriate choice of metrics. Recall that D_u is the subset of functions x in $D \equiv D([0, \infty), \mathbb{R})$ that are unbounded above and have $x(0) \geq 0$. For positive t_1, t_2 , let $D_u(t_1, t_2)$ be the subset of x in D_u with $x^\uparrow(t_1) \geq t_2$. Clearly $D_u(t_1, t_2)$ is a closed subset of D_u . Moreover,

$$D_u = \cap_{m=1}^\infty \cup_{k=1}^\infty D(k, m). \quad (7.7)$$

We now show that the inverse map from $D_u(t_1, t_2) \subseteq D_u([0, t_1], \mathbb{R}, M_1)$ to $D([0, t_2], \mathbb{R}, M_1'')$ is Lipschitz.

Lemma 9.7.4. For $t > 0$, let d_t'' be the M_1'' metric on $D([0, t], \mathbb{R})$ and let d_t be the M_1 metric on $D([0, t], \mathbb{R})$. If $x_1, x_2 \in D_u(t_1, t_2)$, then

$$d_{t_2}''(x_1^{-1}, x_2^{-1}) \leq d_{t_1}(x_1^\uparrow \wedge t_2, x_2^\uparrow \wedge t_2) \leq d_{t_1}(x_1^\uparrow, x_2^\uparrow) \leq d_{t_1}(x_1, x_2) . \quad (7.8)$$

where $(x_i^\uparrow \wedge t_2)(s) = x_i^\uparrow(s) \wedge t_2$, $0 \leq s \leq t_1$.

Proof. For $x_i \in D_u(t_1, t_2)$, let (u_i, r_i) be an arbitrary M_1 parametric representation of $x_i^\uparrow \wedge t_2$ over $[0, t_1]$. Then (r_i, u_i) is an M_1'' parametric representation of x_i^{-1} over $[0, t_2]$ with the special property that $u_i(1) = t_1$. Hence

$$d_{t_2}''(x_1^{-1}, x_2^{-1}) \leq d_{t_1}(x_1^\uparrow \wedge t_2, x_2^\uparrow \wedge t_2) . \quad (7.9)$$

It is not difficult to see that

$$d_{t_1}(x_1^\uparrow \wedge t_1, x_2^\uparrow \wedge t_2) \leq d_{t_1}(x_1^\uparrow, x_2^\uparrow) \leq d_{t_1}(x_1, x_2) .$$

Hence the proof is complete. ■

Lemmas 9.7.2 and 9.7.4 imply that the inverse map from $D_u([0, \infty), \mathbb{R}, M_1)$ to $D_u([0, \infty), \mathbb{R}, M_1')$ is continuous, which is weaker than Theorem 13.6.2 in the book. We now want to establish an analog of Theorem 9.2.3. For that purpose, we need both x and x^{-1} to be Lipschitz on $[0, t]$ for all $t > 0$. The following lemma provides natural conditions.

Lemma 9.7.5. (conditions for both x and x^{-1} to be Lipschitz) If $x \in D_u$ is absolutely continuous, i.e., $x(t) = \int_0^t \dot{x}(s)ds$ for $t > 0$, with $\dot{x} \in D$ and with $l(t) \leq \dot{x}(t) \leq u(t)$ for all $t \geq 0$ where $0 < l^\downarrow(t) < u^\uparrow(t) < \infty$ for all t , then

$$x^{-1}(t) = \int_0^t [1/\dot{x}(x^{-1}(s))]ds \quad \text{for all } t > 0 \quad (7.10)$$

and x and x^{-1} are both Lipschitz on $[0, t]$ for all $t > 0$, with

$$\dot{x}^{-1}(t) \equiv \frac{d}{dt}(x^{-1})(t) = 1/\dot{x}(x^{-1}(t)) . \quad (7.11)$$

Proof. Clearly x is strictly increasing and continuous, so that x is a homeomorphism of $[0, \infty)$ and $x \circ x^{-1} = e$, where $\circ$ is the composition map. Thus

$$x(x^{-1}(t)) = \int_0^{x^{-1}(t)} \dot{x}(s)ds = t, \quad t \geq 0 ,$$

which implies that

$$x^{-1}(t) = \int_0^t [1/\dot{x}(x^{-1}(s))]ds, \quad t \geq 0.$$

The Lipschitz properties hold because

$$|x(t_2) - x(t_1)| = \left| \int_{t_1}^{t_2} \dot{x}(s)ds \right| \leq u^\uparrow(t_2)|t_2 - t_1|$$

and

$$|x^{-1}(t_2) - x^{-1}(t_1)| = \left| \int_{t_1}^{t_2} [1/\dot{x}(x^{-1}(s))]ds \right| \leq |t_2 - t_1|/l^\downarrow(t_2). \quad \blacksquare$$

We now want to establish an analog of Theorem 9.2.3. Since the M_1'' analog of Lemma 9.2.2 is evident, we only establish the M_1'' analog of Lemma 9.2.1.

Lemma 9.7.6. (reduction of convergence to the derivative with the M_1'' topology) *Suppose that x is Lipschitz on $[0, t]$ with Lipschitz constant K . Let d_t'' be the M_1'' metric on $D([0, t], \mathbb{R})$. Then*

$$d_t''(x_1 - x, x_2 - x) \leq (1 + K)d_t''(x_1, x_2).$$

Proof. For all $\epsilon > 0$, there exists $\eta(\epsilon) > 0$ and parametric representations $(u_{i,\epsilon}, r_{i,\epsilon}) \in \Pi_t''(x_i)$ such that

$$\|u_{1,\epsilon} - u_{2,\epsilon}\| \vee \|r_{1,\epsilon} - r_{2,\epsilon}\| \leq (1 + \eta(\epsilon))d_t''(x_1, x_2). \quad (7.12)$$

We now want natural modifications of the parametric representations of x_i to serve as parametric representations of x and $x_i - x$. To obtain such parametric representations for x , we need to allow for the line segment joining $(x(0), 0)$ to $(0, 0)$. Hence we first modify the parametric representations of x_i . Let $(u'_{i,\epsilon}, r'_{i,\epsilon}) \in \Pi_t''(x_i)$ be scaled versions of the parametric representations $(u_{i,\epsilon}, r_{i,\epsilon})$ on $[\delta, 1]$ with $(u'_{i,\epsilon}(s), r'_{i,\epsilon}(s)) = (0, 0)$, $0 \leq s \leq \delta$, i.e.,

$$(u'_{i,\epsilon}(\delta + s), r'_{i,\epsilon}(\delta + s)) = (u_{i,\epsilon}((1 - \delta)^{-1}s), r_{i,\epsilon}((1 - \delta)^{-1}s)), \quad 0 \leq s \leq 1 - \delta. \quad (7.13)$$

Then

$$\|u'_{1,\epsilon} - u'_{2,\epsilon}\| \vee \|r'_{1,\epsilon} - r'_{2,\epsilon}\| = \|u_{1,\epsilon} - u_{2,\epsilon}\| \vee \|r_{1,\epsilon} - r_{2,\epsilon}\|. \quad (7.14)$$

Since $x \in C$, $(u''_{i,\epsilon}, r'_{i,\epsilon}) \in \Pi''(x)$ for $i = 1, 2$, if

$$u''_{i,\epsilon}(s) = \begin{cases} x \circ r'_{i,\epsilon}, & \delta \leq s \leq 1 \\ 0, & s = 0 \end{cases}$$

with $u''_{i,\epsilon}$ defined by linear interpolation on $(0, \delta)$. Then $(u'_{i,\epsilon} - u''_{i,\epsilon}, r'_{i,\epsilon}) \in \Pi''(x_i - x)$ and

$$\begin{aligned} d''_t(x_1 - x, x_2 - x) &\leq \|(u'_{1,\epsilon} - u''_{1,\epsilon}) - (u'_{2,\epsilon} - u''_{2,\epsilon})\| \vee \|r'_{1,\epsilon} - r'_{2,\epsilon}\| \\ &\leq (\|u'_{1,\epsilon} - u'_{2,\epsilon}\| + \|x \circ r'_{1,\epsilon} - x \circ r'_{2,\epsilon}\|) \vee \|r'_{1,\epsilon} - r'_{2,\epsilon}\| \\ &\leq (1 + K)(1 + \eta(\epsilon))d''_t(x_1, x_2). \end{aligned}$$

Since $\eta(\epsilon) \rightarrow 0$ as $\epsilon \rightarrow 0$, the proof is complete. ■

We now obtain the M'_1 -analog of Theorem 9.2.3. By Lemma 9.7.2, the M'_1 and M''_1 topologies agree on $D([0, \infty), \mathbb{R})$.

Theorem 9.7.1. *Suppose that $x, x_\epsilon \in D_u([0, \infty), \mathbb{R})$ and that x satisfies the condition of Lemma 9.7.5. If $d_t(x_\epsilon - x, \epsilon y) = o(\epsilon)$ as $\epsilon \rightarrow 0$ for t in a sequence $\{t_k\}$ with $t_k \rightarrow \infty$ as $k \rightarrow \infty$, for which a sufficient condition is $\|\epsilon^{-1}(x_\epsilon - x) - y\|_t \rightarrow 0$ as $\epsilon \downarrow 0$ for all $t > 0$, then*

$$d'(\epsilon^{-1}[x_\epsilon^{-1} - x^{-1}], \epsilon^{-1}[(x + \epsilon y)^{-1} - x^{-1}]) \rightarrow 0 \quad \text{as } \epsilon \downarrow 0, \quad (7.15)$$

where d' is the M'_1 metric on $D([0, \infty), \mathbb{R})$.

Proof. For any $t_2 > 0$, choose t_1 such that $d_{t_1}(x_\epsilon - x, \epsilon y) = o(\epsilon)$ as $\epsilon \downarrow 0$ and $x^{-1}(t_2) < t_1$. The assumptions imply that $\|x_\epsilon - x\|_{t_1} \rightarrow 0$ and $\|\epsilon y\|_{t_1} \rightarrow 0$ as $\epsilon \downarrow 0$. Hence, for all sufficiently small ϵ , $x_\epsilon, x + \epsilon y \in D_u(t_1, t_2)$. On $D_u(t_1, t_2)$, we can apply Lemmas 9.7.4 and 9.7.6, and the M''_1 analog of Lemma 9.2.2 to conclude for $\epsilon \leq 1$ that there are constants K_1 and K_2 such that

$$\begin{aligned} &d''_{t_2}(\epsilon^{-1}[x_\epsilon^{-1} - x^{-1}], \epsilon^{-1}[(x + \epsilon y)^{-1} - x^{-1}]) \\ &\leq \epsilon^{-1}d''_{t_2}(x_\epsilon^{-1} - x^{-1}, (x + \epsilon y)^{-1} - x^{-1}) \\ &\leq K_1\epsilon^{-1}d''_{t_2}(x_\epsilon^{-1}, (x + \epsilon y)^{-1}) \\ &\leq K_1\epsilon^{-1}d_{t_1}(x_\epsilon, x + \epsilon y) \\ &\leq K_1K_2\epsilon^{-1}d_{t_1}(x_\epsilon - x, \epsilon y) \\ &\leq K_1K_2\|(\epsilon^{-1}(x_\epsilon - x) - y)\|_{t_1}. \end{aligned} \quad (7.16)$$

This argument applies for arbitrarily large t_2 provided that we increase t_1 appropriately. ■

We now focus on the derivative of the inverse map. Let

$$z_\epsilon \equiv z_\epsilon(x, y) \equiv \epsilon^{-1}[(x + \epsilon y)^{-1} - x^{-1}] . \quad (7.17)$$

We first observe that z_ϵ in (7.17) is monotone decreasing in y .

Lemma 9.7.7. *For any $x \in D_u$ and $y \in D$, if $y_1(t) \leq y_2(t)$ for all t , then $z_\epsilon(x, y_1)(t) \geq z_\epsilon(x, y_2)(t)$ for all ϵ and t , where z_ϵ is defined in (7.17).*

We now show that it suffices to consider piecewise-constant functions y , because under regularity conditions, $z_\epsilon(x, y)$ as a function of y is Lipschitz. Hence, for x and y given, we can replace y by $y_c \in D_c$.

Lemma 9.7.8. *Suppose that $x \in D_u$, $\dot{x} \in D$, $y_1 \in D$, $t_1 = x^{-1}(t_2) + 1$, $0 < a \leq \|\dot{x}\|_{t_1} < \infty$ and $\|y_1\|_{t_1} \leq K$. If $\|y_1 - y_2\|_{t_1} < 1$, then*

$$\|z_\epsilon(x, y_1) - z_\epsilon(x, y_2)\|_{t_2} \leq (2/a)\|y_1 - y_2\|_{t_1}$$

provided that $\epsilon \leq a/[K + 1]$.

Proof. By the monotonicity established in Lemma 9.7.7,

$$z_\epsilon(x, y_1 - \delta) \geq z_\epsilon(x, y_1), z_\epsilon(x, y_2) \geq z_\epsilon(x, y_1 + \delta)$$

on $[0, t]$ provided that $\|y_1 - y_2\|_{t_1} \leq \delta$ for a suitably large t_1 . For the given t_1 and $\delta \leq 1$,

$$\begin{aligned} (x + \epsilon y_1)^{-1}(t) &\leq (x + \epsilon(y_1 - \delta))^{-1}(t) \\ &\leq (x - \epsilon(K + \delta))^{-1}(t) \\ &\leq x^{-1}(t) + \frac{\epsilon(K + \delta)}{a} \leq t_1 \end{aligned}$$

provided that $\delta \leq 1$ and $\epsilon \leq a/(K + 1)$. Hence, if $\|\dot{x}\|_{t_1} \geq a$ and $\|y_1\|_{t_1} \leq K$ for that t_1 , the inverses are all contained in $[0, t_1]$. Then, for $\|y_1 - y_2\|_{t_1} \leq \delta \leq 1$,

$$\begin{aligned} \|z_\epsilon(x, y_1) - z_\epsilon(x, y_2)\|_{t_2} &\leq \|z_\epsilon(x, y_1 + \delta) - z_\epsilon(x, y_1 - \delta)\|_{t_2} \\ &= \epsilon^{-1}\|(x + \epsilon(y_1 - \delta))^{-1} - (x + \epsilon(y_1 + \delta))^{-1}\|_{t_2} \\ &\leq x^{-1}(t_2) + 2\delta/a . \quad \blacksquare \end{aligned}$$

We now establish pointwise convergence. For this purpose, let

$$Pos(x) = \{t \geq 0 : x(t) > 0\} . \quad (7.18)$$

We obtain the following result by examining the indicated cases.

Theorem 9.7.2. *If $y \in D$ and $x \in D_u$ satisfies the condition of Lemma 9.7.5, then*

$$z_\epsilon(t) \equiv \epsilon^{-1}[(x + \epsilon y)^{-1}(t) - x^{-1}(t)] \rightarrow z(t) \quad \text{in } \mathbb{R} \quad \text{as } \epsilon \downarrow 0$$

for each t , where

$$(i) \quad z(t) = \frac{-y(x^{-1}(t)-)}{\dot{x}(x^{-1}(t)-)} < 0 \quad (7.19)$$

if $y(x^{-1}(t)-) > 0$;

$$(ii) \quad z(t) = \frac{-y(x^{-1}(t))}{\dot{x}(x^{-1}(t))} > 0 \quad (7.20)$$

if $y(x^{-1}(t)-) < 0$ and $y(x^{-1}(t)) < 0$ or if $y(x^{-1}(t)-) = 0$, $\sup\{Pos(y \circ x^{-1}) \cap [0, t)\} < t$ and $y(x^{-1}(t)) < 0$;

$$(iii) \quad z(t) = 0 \quad (7.21)$$

otherwise: if one of: (a) $y(x^{-1}(t)-) = 0$ and $\sup\{Pos(y \circ x) \cap [0, t)\} = t$, (b) $y(x^{-1}(t)-) < 0$ and $y(x^{-1}(t)) = 0$, (c) $y(x^{-1}(t)-) = 0$, $\sup\{Pos(y \circ x) \cap [0, t)\} < t$ and $y(x^{-1}(t)) = 0$, or (d) $y(x^{-1}(t)-) < 0 < y(x^{-1}(t))$.

Consequently, z is either left-continuous or right-continuous at t unless $y(x(t)-) < 0 < y(x(t))$, in which case $z(t-) > z(t) > z(t+)$.

Proof. It is elementary that $z_\epsilon(t)$ converges pointwise to $z(t)$ for $z(t)$ in (7.20) when both $\dot{x}$ and y are continuous at $x^{-1}(t)$, so that z is continuous at t . For the other cases, we apply Lemma 9.7.8 to approximate y by a piecewise-constant function. We then exploit Lemma 9.7.7 and the fact that $\dot{x}$ and y are elements of D . We obtain the conclusions by examining the different cases. ■

Remark 9.7.1. In order to have the pointwise convergence in Theorem 9.7.2, at a single t , it suffices to have the conditions on x and y hold only in a neighborhood of $x^{-1}(t)$. Then x need not be absolutely continuous or strictly increasing everywhere.

Remark 9.7.2. We have difficulty at some t if x is only an increasing homeomorphism of $[0, \infty)$. Then we can have $\dot{x}(x^{-1}(t)) = 0$ and $\dot{x}^{-1}(t) = \infty$ for some t , so that $z_\epsilon(t) \rightarrow \infty$ as $\epsilon \downarrow 0$.

We now want to establish M'_1 convergence in D . However, first we note that the limit z does not necessarily belong to D , because it may be neither left-continuous nor right-continuous at discontinuity points.

Example 9.7.1. *We need not have $z \in D$. To see that we need not have $z \in D$, even if $\dot{x} \in C$, let $x = e$ and let $y = -I_{[0,1)} + I_{[1,\infty)}$. Then*

$$z_\epsilon(t) = I_{[0,1-\epsilon)}(t) + \epsilon^{-1}(1-t)I_{[1-\epsilon,1+\epsilon)}(t) - I_{[1+\epsilon,\infty)}(t) \quad (7.22)$$

and

$$z = I_{[0,1)} - I_{(1,\infty)} \quad (7.23)$$

so that $z(1) = 0$, but $z(1-) = 1$ and $z(1+) = -1$. However, $z(1)$ is in between $z(1-)$ and $z(1+)$. ■

Since $z(t)$ lies between $z(t-)$ and $z(t+)$ for all t , the space D^* of such functions with the M_1 and M' topologies is equivalent to D because functions in D and D^* have the same graphs.

Theorem 9.7.3. (conditions for convergence to the right-continuous version) *If $y \in D$ and $x \in D_u$ satisfies the condition of Lemma 9.7.5 with $\dot{x} \in D$, then*

$$z_\epsilon \rightarrow z_+ \quad \text{in } (D, M'_1) \quad \text{as } \epsilon \downarrow 0$$

for z_ϵ in (7.17) and z_+ the right-continuous version of z , i.e. $z_+(t) = z(t+)$, $t \geq 0$ and z in (7.19). If $z_+(0) = 0$, the convergence is in M_1 .

Proof. First, for x and y given, with $\dot{x}$ satisfying the conditions of Lemma 9.7.5, the conditions of Lemma 9.7.8 are satisfied. Since $\dot{x} \in D$ and $y \in D$, $z \in D^*$ for z defined in (7.19). Start by replacing z by its right-continuous version, which has the same graph. Invoking Lemma 9.4.1, for any $t > 0$, let $\tilde{z} \in D_c$ be such that $\|z - \tilde{z}\|_t \leq \delta_1$. Suppose that $x^{-1}(t_1)$ and $x^{-1}(t_2)$ are two successive discontinuity points of y (where $t_1, t_2 < t$), regarded as an element of D_c . Suppose that $y(s) = c > 0$ in $[x^{-1}(t_1), x^{-1}(t_2))$. Then, for any $\delta_2 > 0$, $z_\epsilon(s) \uparrow z(s)$ in $(t_1 + \delta_2, t_2 - \delta_2)$. Since z_ϵ and $\tilde{z} + \delta_1$ are both continuous in $(t_1 + \delta_2, t_2 - \delta_2)$, we can apply Dini's theorem to conclude that $z_\epsilon(s) \wedge \tilde{z}(s) - \delta_1$ converges uniformly to $\tilde{z}(s) - \delta_1$ in $(t_1 + \delta_2, t_2 - \delta_2)$. Similarly, if $y(s) = c < 0$ in $[t_1, t_2)$, then we can conclude that $z_\epsilon(s) \vee (\tilde{z}(s) + \delta_1)$ converges uniformly to $\tilde{z}(s) + \delta_1$ in $(t_1 + \delta_2, t_2 - \delta_2)$. It thus suffices to establish local M_1 convergence at each of the isolated discontinuity points of $\tilde{z}$; see Theorem ???. However, z_ϵ is monotone in a neighborhood of each of these discontinuity points for all sufficiently small ϵ . Together with the pointwise convergence at all continuity points established in Theorem 9.7.2, this implies the required local M_1 convergence. To get the strengthened convergence to M_1 , apply Theorem 13.6.3 in the book. ■

The derivative result in Theorem 9.7.3 holds for arbitrary $y \in D$. By applying Theorem 9.7.1, we obtain a corresponding preservation result, but only under the extra condition of uniform convergence of $\epsilon^{-1}(x_\epsilon - x)$ to y as $\epsilon \downarrow 0$, which holds if $y \in C$.

Below let U be the topology on $D([0, \infty), \mathbb{R})$ of uniform convergence over compact subsets.

Corollary 9.7.1. *Suppose that $x_\epsilon, x \in E$. Under the conditions of Theorem 9.7.3, if $\|\epsilon^{-1}(x_\epsilon - x) - y\|_t \rightarrow 0$ as $\epsilon \downarrow 0$ for all $t > 0$, then*

$$\epsilon^{-1}(x_\epsilon^{-1} - x^{-1}) \rightarrow z_+ \quad \text{in } (D, M_1^t) \quad \text{as } \epsilon \downarrow 0$$

for z_+ as in Theorem 9.7.3.

9.8. Chapter Notes

As indicated at the outset, this chapter is largely based on Mandelbaum and Massey (1995). They formulate convergence preservation in terms of the directional derivative. We focus on the second term of the triangle inequality in (2.1). Thus The results in Section 9.2 here are new. It would be nice if the upper bound $K\epsilon^{-1}d_1(x_\epsilon - x, y)$ in Theorem 9.2.3 could be replaced by $Kd_1(\epsilon^{-1}(x_\epsilon - x), y)$ under reasonable regularity conditions. The existing bound in terms of $K\epsilon^{-1}d_1(x_\epsilon - x, y)$ may be suitable for applying strong approximations. It thus also would be nice to develop such strong approximations to apply with Theorem 9.2.3 here.

Section 9.3 on the derivative of the supremum function is also based on Mandelbaum and Massey (1995). We provide extensions allowing the functions x and y appearing in $z_\epsilon(x, y)$ in (3.2) to be discontinuous. We also do not require that the limit z have only finitely many discontinuities in each finite interval. The arguments are quite a bit more complicated as a result. Some simplification is achieved here by exploiting approximations by piecewise-constant functions. In particular, for establishing M_1 convergence, Lemma 9.4.2 is key.

Given the intimate connection between the reflection and supremum maps, most of the work on the derivative of the reflection map in Section 9.5 is done in Sections 9.3 and 9.4. The application of Sections 9.3 – 9.5 in Section 9.6 to obtain heavy-traffic limits for nonstationary queues also follows Mandelbaum and Massey (1995). They focused on the $M_t/M_t/1$ queue with fixed arrival-rate and service-rate functions $\lambda^+(t)$ and $\lambda^-(t)$, drawing on the strong approximation for Poisson processes. We show how

the results can be generalized by applying convergence-preservation results for the composition function with nonlinear centering in Chapter 13 of the book.

Section 9.7 on the derivative of the inverse function is new. The M_1'' topology extends the M_1' topology introduced in Puhalskii and Whitt (1997).

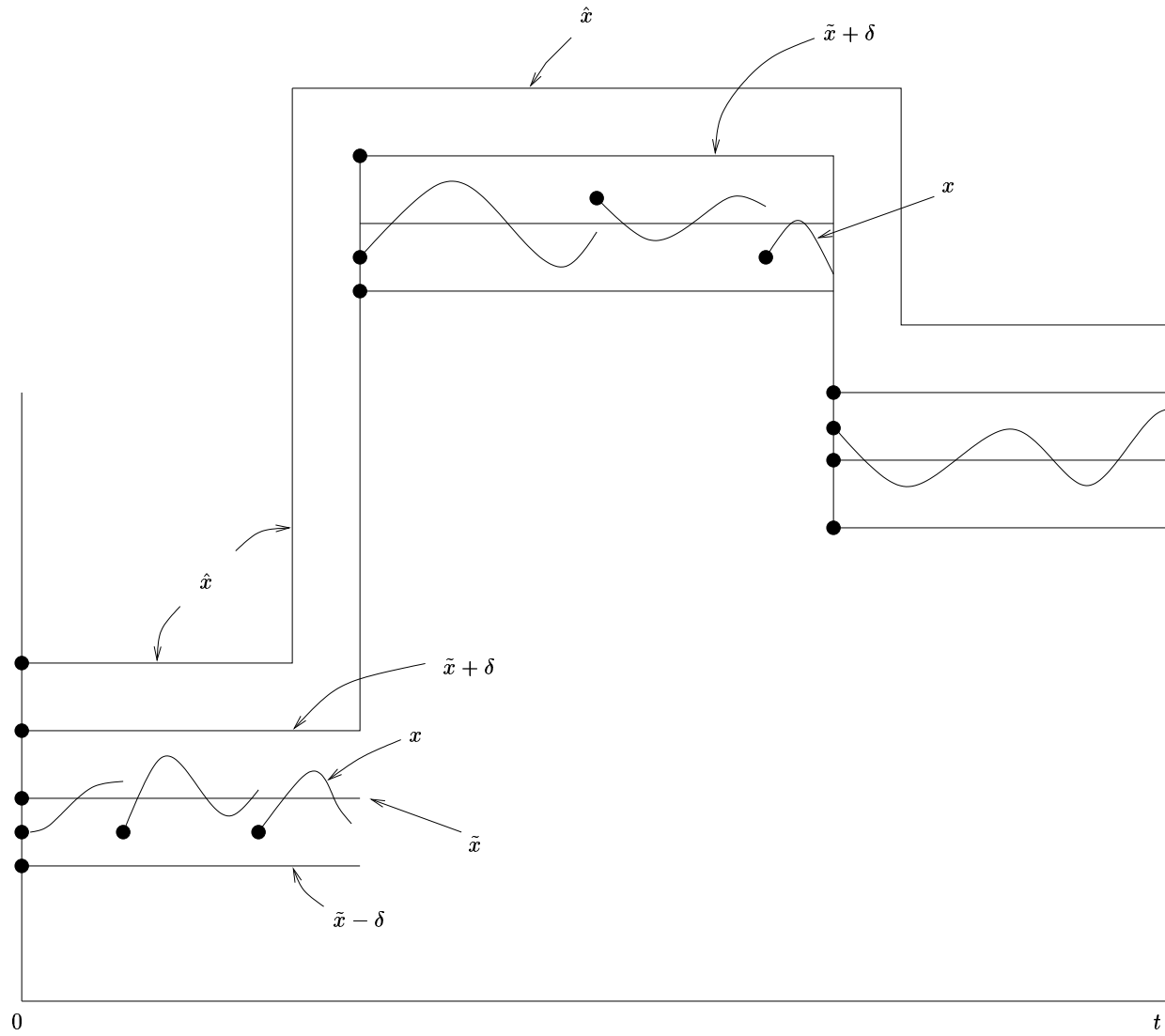


Figure 9.1: A possible function x , piecewise-constant approximation $\tilde{x}$, upper bound $\tilde{x} + \delta$ and upper boundary $\hat{x}$ of the δ -neighborhood of the graph $\Gamma_{\tilde{x}+\delta}$ used in the proof of Theorem 9.4.1.

Figure 9.2: Graphs of the time-dependent arrival-rate and service-rate functions $(\lambda^+(t), \lambda^-(t))$ with λ^- constant, the functions $(-x, (-x)^\dagger)$, the set-valued function Φ_{-x} and the limits q and Q for a typical realization of the $M_t/M_t/1$ queue.

Chapter 10

Errors Discovered in the Book

This chapter contains a list of all errors in the book found since publication.

Chapter 11

Bibliography

Bibliography

- [1] Abate, J., Choudhury, G. L. and Whitt, W. (1999) An introduction to numerical transform inversion and its application to probability models. *Computational Probability*, W. Grassman (ed.), Kluwer, Boston, 257–323.
- [2] Abate, J. and Whitt, W. (1992a) The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems* 10, 5–88.
- [3] Abate, J. and Whitt, W. (1992b) Numerical inversion of probability generating functions. *Operations Res. Letters* 12, 245–251.
- [4] Abate, J. and Whitt, W. (1995) Numerical inversion of Laplace transforms of probability distributions. *ORSA J. Computing* 7, 36–43.
- [5] Abate, J. and Whitt, W. (1996) An operational calculus for probability distributions via Laplace transforms. *Adv. Appl. Prob.* 28, 75–113.
- [6] Albert, A. (1966) Fixed size confidence ellipsoids for linear regression parameters. *Ann. Math. Statist.* 37, 1602–1630.
- [7] Aldous, D. (1978) Stopping times and tightness. *Ann. Prob.* 6, 335–340.
- [8] Anscombe, F. J. (1952) Large sample theory for sequential estimation. *Proc. Cambridge Philos. Soc.* 48, 600–607.
- [9] Anscombe, F. J. (1953) Sequential estimation. *J. Roy. Statist. Soc. Ser. B* 15, 1–21.
- [10] Asmussen, S. (1987) *Applied Probability and Queues*, Wiley, New York.
- [11] Baccelli, F. and Brémaud, P. (1994) *Elements of Queueing Theory*, Springer, New York.

- [12] Beran, J. (1994) *Statistics for Long-Memory Processes*, Chapman and Hall, New York.
- [13] Bertoin, J. (1996) *Lévy Processes*, Cambridge University Press, Cambridge, UK.
- [14] Billingsley, P. (1968) *Convergence of Probability Measures*, Wiley, New York.
- [15] Billingsley, P. (1999) *Convergence of Probability Measures*, second edition, Wiley, New York.
- [16] Bingham, N. H. (1975) Fluctuation theory in continuous time. *Adv. Appl. Prob.* 7, 705–766.
- [17] Bingham, N. H., Goldie, C. M. and Teugels, J. L. (1989) *Regular Variation*, Cambridge University Press, Cambridge, UK.
- [18] Bondesson, L. (1992) *Generalized Gamma Convolutions and Related Classes of Distributions and Densities*, Springer, New York.
- [19] Box, G. E. P. and Jenkins, G. M. (1970) *Time Series Analysis: Forecasting and Control*, Holden Day, San Francisco.
- [20] Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1994) *Time Series Analysis, Forecasting and Control*, third edition, Prentice Hall, NJ.
- [21] Bratley, P., Fox, B. L. and Schrage, L. E. (1987) *A Guide to Simulation*, second edition, Springer, New York.
- [22] Chen, H. and Mandelbaum, A. (1994) Hierarchical modeling of stochastic networks, part II: strong approximations. *Stochastic Modeling and Analysis of Manufacturing Systems*, D. D. Yao, ed., Springer-Verlag, New York, 107–131.
- [23] Chen, H. and Shen, X. (2000) Strong approximations for multiclass feedforward queueing networks. *Ann. Appl. Prob.* 10, 828–876.
- [24] Chen, H. and Yao, D. D. (2001) *Fundamentals of Queueing Networks: Performance, Asymptotics and Optimization*, Springer, New York.
- [25] Choudhury, G. L., Lucantoni, D. M. and Whitt, W. (1994) Multi-dimensional transform inversion with application to the transient $M/G/1$ queue. *Ann. Appl. Prob.* 4, 719–740.

- [26] Chow, Y. S. and Robbins, H. (1965) On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *Ann. Math. Statist.* 36, 457–462.
- [27] Crane, M. A. and Lemoine, A. J. (1977) *An Introduction to the Regenerative Method for Simulation Analysis. Lecture Notes in Control and Inform. Sci.*, Springer, New York.
- [28] Csörgő, M. and Horvath, L. (1993) *Weighted Approximations in Probability and Statistics*, Wiley, New York.
- [29] Csörgő, M. and Révész, P. (1981) *Strong Approximations in Probability and Statistics*, Academic Press, New York.
- [30] Damerddji, H. (1994) Strong consistency of the variance estimator in steady-state simulation output analysis. *Math. Oper. Res.* 19, 494–512.
- [31] Damerddji, H. (1995) Mean-square consistency of the variance estimator in steady-state simulation output analysis. *Operations Res.* 43, 282–291.
- [32] Dudley, R. M. (1968) Distances of probability measures and random variables. *Ann. Math. Statist.* 39, 1563–1572.
- [33] Dunford, N. and Schwartz, J. T. (1958) *Linear Operators. Part I: General Theory*. Interscience, New York.
- [34] Einmahl, U. (1989) Extensions of results of Komlós, Major and Tusnády to the multivariate case. *J. Multivariate Anal.* 28, 20–68.
- [35] El-Taha, M. and Stidham, S., Jr. (1999) *Sample-Path Analysis of Queueing Systems*, Kluwer, Boston.
- [36] Ethier, S. N. and Kurtz, T. G. (1986) *Markov Processes, Characterization and Convergence*, Wiley, New York.
- [37] Feller, W. (1968) *An Introduction to Probability Theory and its Applications*, vol. I, third edition, Wiley, New York.
- [38] Feller, W. (1971) *An Introduction to Probability Theory and its Applications*, vol. II, second edition, Wiley, New York.
- [39] Fishman, G. S. (1977) Achieving specific accuracy in simulation output analysis. *Comm. ACM* 20, 310–315.

- [40] Fox, B. L. and Glynn, P. W. (1989) Replication schemes for limiting expectations. *Prob. Eng. Inf. Sci.* 3, 299–318.
- [41] Gleser, L. J. (1965) On the asymptotic theory of fixed-size confidence bounds for linear regression parameters. *Ann. Math. Statist.* 36, 463–467. [Correction note: 37 (1966) 1053–1055].
- [42] Glynn, P. W. (1994) Poisson’s equation for the recurrent M/G/1 queue. *Adv. Appl. Prob.* 26, 1044–1062.
- [43] Glynn, P. W. and Heidelberger, P. (1989) Jackknifing under a budget constraint. Unpublished manuscript, Dept. Operations Research, Stanford Univ.
- [44] Glynn, P. W. and Iglehart, D. L. (1988) A new class of strongly consistent variance estimators for a steady-state simulation. *Stochastic Process. Appl.* 16, 71–80.
- [45] Glynn, P. W. and Meyn, S. (1996) A Liapounov bound for solutions of the Poisson equation. *Ann. Probab.* 24, 916–931.
- [46] Glynn, P. W. and Whitt, W. (1987) Sufficient conditions for functional-limit-theorem versions of $L = \lambda W$. *Queueing Systems* 1, 279–287.
- [47] Glynn, P. W. and Whitt, W. (1988) Ordinary CLT and WLLN versions of $L = \lambda W$. *Math. Oper. Res.* 13, 674–692.
- [48] Glynn, P. W. and Whitt, W. (1989) Indirect estimation via $L = \lambda W$. *Operations Res.* 37, 82–103.
- [49] Glynn, P. W. and Whitt, W. (1991a) A new view of the heavy-traffic limit theorem for the infinite-server queue. *Adv. Appl. Prob.* 23, 188–209.
- [50] Glynn, P. W. and Whitt, W. (1991b) Departures from many queues in series. *Ann. Appl. Prob.* 1, 546–572.
- [51] Glynn, P. W. and Whitt, W. (1992a) The asymptotic validity of sequential stopping rules for stochastic simulations. *Ann. Appl. Prob.* 2, 180–198.
- [52] Glynn, P. W. and Whitt, W. (1992b). The asymptotic efficiency of simulation estimators. *Operations Res.* 40, 505–520.

- [53] Glynn, P. W. and Whitt, W. (1993) Limit theorems for cumulative processes. *Stoch. Proc. Appl.* 47, 299–314.
- [54] Glynn, P. W. and Whitt, W. (2000) N&S conditions for the Markov chain CLT. AT&T Labs, Florham Park, NJ.
- [55] Gnedenko, B. V. and Kolmogorov, A. N. (1968) *Limit Distributions for Sums of Independent Random Variables*, revised edition, Addison Wesley, Reading, MA.
- [56] Govindarajulu, Z. (1987) *The Sequential Statistical Analysis of Hypothesis Testing, Point and Interval Estimation, and Decision Theory*, 2nd ed. American Sciences Press, Columbus, Ohio.
- [57] Granger, C. W. J. and Joyeux, R. (1980) An introduction to long-range time series models and fractional differencing. *J. Time Ser. Anal.* 1, 15–30.
- [58] Gut, A. (1988) *Stopped Random Walks*, Springer-Verlag, New York.
- [59] Halmos, P. R. (1956) *Measure Theory* Van Nostrand, Princeton, NJ.
- [60] Heidelberger, P. and Welch, P. D. (1981a) A spectral method for confidence interval generation and run length control in simulations. *Comm. ACM* 24, 233–245.
- [61] Heidelberger, P. and Welch, P. D. (1981b) Adaptive spectral methods for simulation output analysis. *IBM J. Res. Develop.* 25, 860–876.
- [62] Heidelberger, P. and Welch, P. D. (1983) Simulation run length control in the presence of an initial transient. *Oper. Res.* 31, 1109–1144.
- [63] Hosking, J. R. M. (1981) Fractional differencing. *Biometrika* 68, 165–176.
- [64] Jacod, J. and Shiryaev, A. N. (1987) *Limit Theorems for Stochastic Processes*, Springer-Verlag, New York.
- [65] Kella, O. (1993) Parallel and tandem fluid networks with dependent Lévy inputs. *Ann. Appl. Prob.* 3, 682–695.
- [66] Kella, O. (1996) Stability and non-product form of stochastic fluid networks with Lévy inputs. *Ann. Appl. Prob.* 6, 186–199.

- [67] Kella, O. and Whitt, W. (1992a) A tandem fluid network with Lévy input. In *Queues and Related Models*, I. Basawa and N. V. Bhat (eds.), Oxford University Press, Oxford, 112–128.
- [68] Kella, O. and Whitt, W. (1992b) Useful martingales for stochastic storage processes with Lévy input, *J. Appl. Prob.* 29, 396–403.
- [69] Kella, O. and Whitt, W. (1996) Stability and structural properties of stochastic storage networks. *J. Appl. Prob.* 33, 1169–1180.
- [70] Kemeny, J. G. and Snell, J. L. (1960) *Finite Markov Chains*, Van Nostrand, Princeton.
- [71] Kemeny, J. G. and Snell, J. L. (1961) Finite continuous time Markov chains. *Theor. Probability Appl.* 6, 101–105.
- [72] Kennedy, D. P. (1973) Limit theorems for finite dams. *Stoch. Proc. Appl.* 1, 269–278.
- [73] Kogan, Ya., Liptser, R. Sh. and Smorodinskii, A. V. (1986) Gaussian diffusion approximations of Markovian closed models of computer communication networks. *Problemy Peredachi Informatsii* 22, 49–65.
- [74] Komlós, J., Major, P. and Tusnády, G. (1975) An approximation of partial sums of independent R.V.'s and the sample DF, I. *Zeitschrift für Wahrscheinlichkeitstheorie verw. Gebiete* 32, 111–131.
- [75] Komlós, J., Major, P. and Tusnády, G. (1976) An approximation of partial sums of independent R.V.'s and the sample DF, II. *Zeitschrift für Wahrscheinlichkeitstheorie verw. Gebiete* 34, 33–58.
- [76] Kurtz, T. G. (1996) Limit theorems for workload input models. In *Stochastic Networks: Theory and Applications*, F. P. Kelly, S. Zachary and I. Ziedins (eds.) Oxford University Press, Oxford, U.K., 119–139.
- [77] Kushner, H. J. (2001) *Heavy Traffic Analysis of Controlled and Uncontrolled Queueing and Communication Networks*, Springer, New York.
- [78] Lavenberg, S. S. and Sauer, C. H. (1977) Sequential stopping rules for the regenerative method of simulation. *IBM J. Res. Develop.* 21, 545–558.
- [79] Law, A. M. and Carson, J. S. (1979) A sequential procedure for determining the length of a steady-state simulation. *Oper Res.* 27, 1011–1025.

- [80] Law, A. M. and Kelton, W. D. (1982) Confidence intervals for steady-state simulation. II. A survey of sequential procedures. *Management Sci.* 28, 550–562.
- [81] Law, A. M., Kelton, W. D. and Koenig, L. W. (1981) Relative width sequential confidence intervals for the mean. *Comm. Statist. B — Simulation Comput.* 10, 29–39.
- [82] Lindvall, T. (1973) Weak convergence in the function space $D[0, \infty)$. *J. Appl. Prob.* 10, 109–121.
- [83] Lindvall, T. (1992) *Lectures on the Coupling Method*, Wiley, New York.
- [84] Lucantoni, D. M. (1993) The BMAP/G/1 queue: a tutorial. *Models and Techniques for Performance Evaluation of Computer and Communications Systems*, J. Donatiello and R. Nelson (eds.), Springer, New York, 330–358.
- [85] Maigret, N. (1978) Théorème de limite centrale fonctionnel pour une chaîne de Markov récurrente au sens de Harris et positive. *Ann. Inst. H. Poincaré Sect. B (N. S.)* 14, 425–440.
- [86] Mandelbaum, A. and Massey, W. A. (1995) Strong approximations for time-dependent queues. *Math. Oper. Res.* 20, 33–64.
- [87] Mandelbaum, A., Massey, W. A. and Reiman, M. I. (1998) Strong approximations for Markovian service networks. *Queueing Systems* 30, 149–201.
- [88] Mandelbaum, A., Massey, W. A., Reiman, M. I. and Stolyar, A. (1999) Waiting time asymptotics for time varying multiserver queues with abandonments and retries. Proceedings of 37th Allerton Conference on Communication, Control and Computing, September 1999, 1095–1104.
- [89] Massey, W. A. and Whitt, W. (1994) Unstable asymptotics for nonstationary queues. *Math. Oper. Res.* 19, 267–291.
- [90] McLeish, D. L. (1976) Functional and random central limit theorems for the Robbins-Monro process. *J. Appl. Probab.* 13, 148–154.
- [91] Miller, R. G. (1964) A trustworthy jackknife. *Ann. Math. Statist.* 35, 1594–1605.
- [92] Miller, R. G. (1974) The jackknife — a review. *Biometrika* 61, 1–15.

- [93] Nadas, A. (1969) An extension of a theorem of Chow and Robbins on sequential confidence intervals for the mean. *Ann. Math. Statist.* 40, 667–671.
- [94] Neuts, M. F. (1989) *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, Marcel Dekker, New York.
- [95] Neveu, J. (1965) *Mathematical Foundations of the Calculus of Probability*, Holden-Day, San Francisco.
- [96] Park, K. and Willinger, W. (2000) *Self-Similar Network Traffic and Performance Evaluation*, Wiley, New York.
- [97] Parthasarathy, K. R. (1967) *Probability Measures on Metric Spaces*, Academic, New York.
- [98] Philipp, W. and Stout, W. (1975) *Almost Sure Invariance Principles for Partial Sums of Weakly Dependent Random Variables*, Mem. Amer. Math. Soc. **161**, Providence, RI.
- [99] Prabhu, N. U. (1998) *Stochastic Storage Processes*, second ed. Springer, New York.
- [100] Prohorov, Yu. V. (1956) Convergence of random processes and limit theorems in probability. *Theor. Probability Appl.* 1, 157–214.
- [101] Puhalskii, A. A. (1994) On the invariance principle for the first passage time. *Math. Oper. Res.* 19, 946–954.
- [102] Puhalskii, A. A. and Whitt, W. (1997) Functional large deviation principles for first-passage-time processes. *Ann. Appl. Prob.* 7, 362–381.
- [103] Resnick, S. and Stărică, C. (1997) Smoothing the Hill estimator. *Adv. Appl. Prob.* 29, 271–293.
- [104] Rogers, L. C. G. (2000) Evaluating first-passage probabilities for spectrally one-sided Lévy processes. *J. Appl. Prob.* 37, 1173–1180.
- [105] Ruppert, D. (1982) Almost sure approximations to the Robbins-Monro and Kiefer-Wolfowitz processes with dependent noise. *Ann. Probab.* 10, 178–187.
- [106] Samorodnitsky, G. and Taqqu, M. S. (1994) *Stable Non-Gaussian Random Processes*, Chapman-Hall, New York.

- [107] Siegmund, D. (1985) *Sequential Analysis*. Springer, New York.
- [108] Simmons, G. F. (1963) *Topology and Modern Analysis*, McGraw-Hill, New York.
- [109] Skorohod, A. V. (1956) Limit theorems for stochastic processes. *Theor. Probability Appl.* 1, 261–290. (also republished in *Skorohod's Ideas in Probability Theory*, V. Korolyuk, N. Portenko and H. Syta (eds.), institute of Mathematics of the National Academy of Sciences of the Ukraine, Kyiv, Ukraine, 23–52.)
- [110] Skorohod, A. V. (1957) Limit theorems for stochastic processes with independent increments. *Theor. Probability Appl.* 2, 138–171.
- [111] Skorohod, A. V. (1961) *Studies in the Theory of Random Processes*, Addison-Wesley, Reading, MA.
- [112] Srivastava, M. S. (1967) On fixed-width confidence bounds for regression parameters and the mean vector. *J. Roy. Statist. Soc. Ser. B* 29, 132–140.
- [113] Starr, N. (1966) The performance of a sequential procedure for the fixed-width interval estimation of the mean. *Ann. Math. Statist.* 37, 36–50.
- [114] Strassen, V. (1965) The existence of probability measures with given marginals. *Ann. Math. Statist.* 36, 423–439.
- [115] Takács, L. (1967) *Combinatorial Methods in the Theory of Stochastic Processes*, Wiley, New York.
- [116] Thorin, O. (1977a) On the infinite divisibility of the Pareto distributions. *Scand. Act. J.* 60, 31–40.
- [117] Thorin, O. (1977b) On the infinite divisibility of the lognormal distribution. *Scand. Act. J.* 60, 121–148.
- [118] Venter, J. H. (1967) An extension of the Robbins-Monro procedure. *Ann. Math. Statist.* 38, 181–190.
- [119] Wetherill, G. B. and Glazebrook, K. D. (1986) *Sequential Methods in Statistics*, 3rd ed., Chapman and Hall, London.

- [120] Whitt, W. (1974) Preservation of rates of convergence under mappings. *Zeitschrift für Wahrscheinlichkeitstheorie verw. Gebiete* 29, 39–44.
- [121] Whitt, W. (1980) Some useful functions for functional limit theorems. *Math. Oper. Res.* 5, 67–85.
- [122] Whitt, W. (1991) A review of $L = \lambda W$ and extensions. *Queueing Systems* 9, 235–268. (Correction Note on $L = \lambda W$. *Queueing Systems* 12, 431–432.)
- [123] Whitt, W. (1992a) Asymptotic formulas for Markov processes with applications to simulation. *Operations Res.* 40, 279–291.
- [124] Whitt, W. (1992b) $H = \lambda G$ and the Palm transformation. *Adv. Appl. Prob.* 24, 755–758.
- [125] Whitt, W. (2001) The reflection map with discontinuities. *Math. Oper. Res.*, to appear.
- [126] Wichura, M. J. (1970) On the construction of almost uniformly convergent random variables with given weakly convergent image laws. *Ann. Math. Statist.* 41, 284–291.

Index

Notation by Chapter, 1

Chapter 1, 1

$\Rightarrow$, weak convergence, 1
 $\pi(P_1, P_2)$, Prohorov metric, 1
 $\mathcal{P} \equiv \mathcal{P}(S)$, 1
 A^ϵ , ϵ -open nbhd., 2
 $\mathcal{G}$, set of real-val. fcts., 2
 $\pi_\gamma(P_1, P_2)$, gen. Proh. metric, 3
 A^- , closure of A , 3
 $F(t)$, cdf, 6
 $F^{-1}(t)$, rt.-cont. inverse, 6
 $rad(A)$, radius of A , 8
 ∂A , boundary of A , 8
 δ_s , Dirac measure, 10
 $Z(S)$, finite signed measures, 16
 B^* , adjoint space, 16
 $p(X_1, X_2)$, in-prob. dist., 17

Chapter 2, 23

$\stackrel{d}{=}$, equality in distribution, 23
 Φ , std. normal cdf, 24
 $\phi_K : D \rightarrow D$, ref. map, 26

Chapter 3, 51

$x^\uparrow$, supremum map, 52, 173
 $\Lambda(\mathbb{R}_+)$, homeomorphisms, 51
 $\mathcal{R}(\alpha)$, regularly varying, 52

Chapter 4, 73

Γ , scaling matrix, 75
 $\nabla g(\mu)$, gradient, 88

Chapter 5, 97

$\psi(s) \equiv \log Ee^{-sL(1)}$, Laplace exp.,
 99

ξ_n , 101

γ_n , 101

Chapter 6, 113

$x(t-)$, left limit, 114
 $[a, b]$, standard segment, 117
 $[[a, b]]$, product segment, 117
 Γ_x , thin graph, 117
 G_x , thick graph, 117
 $\rho(\Gamma_x)$, thin range, 117
 $\rho(G_x)$, thick range, 117
 $\Pi_s(x)$, set of strong par. reps., 118
 $\Pi_w(x)$, set of weak par. reps., 118
 $\beta_t : [0, 1] \rightarrow [0, 1]$, 123
 $\mu_s(x_1, x_2)$, SM_2 dist., 144
 $\Pi_{s,2}(x)$, SM_2 par. reps., 148
 $\Pi_{w,2}(x)$, WM_2 par. reps., 148

Chapter 7, 163

$x \circ y$, composition, 164
 $\phi(x) \equiv x + (-x \vee 0)^\uparrow$, one-dim.
 reflection map, 181
 x^{-1} , inverse map, 183

Chapter 8, 195

$\Psi(x)$, feas. regulator set, 197
 $R \equiv (\psi, \phi)$, mult. ref. map, 197
 $\pi \equiv \pi_{x,Q} : D_\uparrow^k \rightarrow D_\uparrow^k$, 199

Chapter 9, 235

$z_\epsilon \equiv z_\epsilon(x, y) \equiv \epsilon^{-1}[(x + \epsilon y)^\uparrow - x^\uparrow]$,
 243

$\Phi_x^L(t)$, 244

$\Phi_x^R(t)$, 244

$\Phi_x(t)$, 244

- Abate, 22, 44, 48
- adaptedness of ref. map, 202
- addition
 - continuity of
 - for M_1 , 135, 136
 - for M_2 , 159
 - measurability of, 136
- adjoint space, 16
- Albert, 87
- analytic method, 22
- Anscombe, 71, 87
- approximation
 - of functions
 - in D by fcts. in D_c , 116
 - in D_s by fcts. in $D_{s,l}$, 208
 - of graphs by finite sets, 119
- ARMA (p, q) , 47
- arrivals see time averages, ASTA, 52
- Asmussen, 36
- ASTA, 52
- asymptotic
 - validity of seq. stop rule, 78
- asymptotic efficiency of simulation
 - estimators, 73
- asymptotic equivalence of
 - cting. and inv. fcts., 192
- asymptotic variance, 30
 - for a birth-and-death pr., 35
 - for the $M/M/1$ queue, 35
- autocovariance function, 34
- autoregressive moving average, ARMA, 47
- $B(x, \epsilon)$, open ball, 5
- $Bad(x)$, set of bad pts., 248
- Baccelli, 71
- backshift operator, 47
- Banach space
 - perspective on “weak”, 16
- batch Markovian arrival pr., 39
- Beran, 47, 50
- Berry-Esseen theorem, 24
- Bertoin, 41, 43, 100
- Billingsley, 19–22, 40, 165, 166
- Bingham, 81, 98
- birth-and-death process, 35
 - asymptotic variance for, 35
- BMAP, batch Mark. arrival pr., 39
- Bondesson, 44
- book
 - Appendix A, 52, 79, 80, 85
 - Chapter 11, 27
 - Chapter 12, v, 113
 - Chapter 13, iv, v, 38, 51, 163, 235, 237, 277
 - Chapter 14, v, 195
 - Chapter 3, iii, 27
 - Chapter 4, iv, 23
 - Chapter 5, v, 97, 196
 - Chapter 7, 23
 - Chapter 8, v, 97
 - Chapter 9, v, 97
 - condition 12.5.4, 177
 - condition 12.5.5, 177
 - Corollary 12.11.2, 213
 - Corollary 12.11.4, 103
 - Corollary 12.11.6, 161, 193
 - Corollary 12.5.1, 174, 212
 - Corollary 13.4.1, 236
 - Corollary 13.7.1, 236
 - Corollary 13.7.2, 236
 - Corollary 14.3.2, 225
 - Corollary 14.3.4, 211
 - Corollary 14.3.5, 216
 - equation 11.5.3, 125
 - equation 11.5.4, 144
 - equation 12.4.1, 168
 - equation 12.4.3, 168, 171

- equation 12.4.4, 170, 171
- equation 12.5.3, 103
- equation 13.4.1, 238
- equation 13.5.1, 238
- equation 14.2.35, 216
- equation 14.2.6, 216
- equation 14.8.6, 227
- equation 2.3.6, 26
- equation 3.3.2, 114, 121, 181
- equation 3.3.4, 118, 181
- equation 4.5.12, 43
- equation 4.5.13, 43
- equation 4.6.13, 50
- equation 4.6.6, 46
- equation 4.7.1, 50
- equation 5.2.5, 181
- Example 1.4.2, 47
- Example 12.10.1, 145
- Example 12.3.1, 121, 214
- Example 14.5.3, 213
- Figure 11.2, 113
- Figure 13.1, 191
- Lemma 12.4.2, 169
- Lemma 12.5.1, 160
- Lemma 13.4.1, 239
- Lemma 13.5.1, 239
- Lemma 13.6.3, 190
- Lemma 13.8.1, 55
- Lemma 14.3.3, 225
- Lemma 14.3.4, 208, 210
- Proposition 13.2.1, 82
- Section 1.3, 7
- Section 1.4, 7, 23
- Section 10.4.4, iv, 73
- Section 11.2, 139
- Section 11.5, 144
- Section 11.6, 114, 221
- Section 12.11, 163
- Section 12.2, 255
- Section 12.4, 166, 168
- Section 12.6, 134, 163
- Section 12.7, 102, 163
- Section 12.9, 144
- Section 13.4, 259
- Section 13.6, 187, 268
- Section 13.7, 62, 102, 190
- Section 13.8, 55, 102
- Section 14.2, 204
- Section 14.6, 196, 218, 233
- Section 14.7, 233
- Section 2.3, 26, 196
- Section 3.2, 25
- Section 3.4, 17
- Section 3.5, 51, 163, 235
- Section 4.3, 23, 41
- Section 4.4, 30, 41, 90
- Section 4.5, 41
- Section 4.6, 23, 50
- Section 4.7, 23, 41, 50
- Section 5.9, iv, 36, 73
- Section 7.3, 190
- Section 8.5, 22
- Theorem 11.3.1, 5, 9
- Theorem 11.3.4, 20, 21, 76
- Theorem 11.3.5, 17
- Theorem 11.4.5, 77, 85
- Theorem 11.5.2, 164
- Theorem 11.5.3, 164
- Theorem 11.6.1, 221, 231
- Theorem 11.6.7, 221
- Theorem 12.11.1 (v), 167
- Theorem 12.11.2 (iv), 167
- Theorem 12.2.2, 175
- Theorem 12.4.1, 169, 172, 177
- Theorem 12.5.1, 103, 105, 177, 255, 258
- Theorem 12.5.1 (v), 167, 212
- Theorem 12.5.2 (iv), 167
- Theorem 12.7.3, 102
- Theorem 12.9.4, 217

- Theorem 13.2.3, 81, 85
- Theorem 13.3.2, 71, 236, 263, 264
- Theorem 13.4.1, 54
- Theorem 13.6.2, 270
- Theorem 13.6.3, 269, 275
- Theorem 13.7.2, 109, 236
- Theorem 13.7.4, v, 105, 108, 236
- Theorem 13.8.2, 62
- Theorem 14.2.4, 225
- Theorem 14.2.5, 214, 216
- Theorem 14.2.9, 195, 216
- Theorem 14.4.1, 211
- Theorem 14.4.2, 213
- Theorem 14.4.2 (a), 215
- Theorem 14.4.3, 215, 216
- Theorem 14.8.1, 226
- Theorem 14.8.3, 227
- Theorem 14.8.6, 226
- Theorem 3.2.1, 1
- Theorem 3.2.2, 6
- Theorem 3.4.2, 25, 26
- Theorem 3.4.3, 19
- Theorem 3.4.4, 20
- Theorem 4.3.2, 86
- Theorem 4.3.5, 87
- Theorem 4.4.2, 31
- Theorem 4.4.4, 40
- Theorem 4.6.1, 50
- Theorem 5.2.1, 80
- Theorem 5.8.2, 99
- Theorem 8.3.1, v, 100
- Theorem 8.5.2, 99
- Theorem A.5, 53
- Borel-Cantelli theorem, 28
- Box, 47, 48
- Bratley, 75, 87
- Bremaud, 71
- Brownian motion
 - fluctuations of, 28
 - modulus of continuity for paths, 28
 - nondifferentiability of paths, 28
- $C_0 \equiv C \cap D_0$, 164
- $C_{\uparrow} \equiv C \cap D_{\uparrow}$, 164
- $C_{\uparrow\uparrow} \equiv C \cap D_{\uparrow\uparrow}$, 164
- $C_m \equiv C \cap D_m$, 164
- $C(S)$, cont. bdd. real-val. fcts. on S , 1
- Carson, 75, 91
- centering
 - for convergence preservation linear, 188
 - in other direction, 180
- central-server model, 110
- characterizations of
 - SM_1 convergence in D
 - main theorem, 128
 - by linear maps, 136
 - by visits to strips, 138
 - SM_2 convergence in D
 - main theorem, 149
 - by linear maps, 159
 - by local extrema, 160
 - WM_1 convergence in D , 131
 - WM_2 convergence in D , 155
 - feasible regulator set, 199
 - local uniform convergence, 126
 - multidimensional reflection map
 - by complementarity, 199
 - by fixed-point property, 199
 - parametric reps., 122
- Chen, 27
- Choudhury, 22
- Chow, 87
- closed queueing network, 110
- CLT equivalence for cting. fcts., 62

- comparison of
 - SJ_1 and SM_1 metrics, 121
 - SM_1 and WM_1 topologies, 121
- complementarity and reflection, 199
- complete
 - metric space, 139
- composition map, 68, 164
 - continuity of, 165
 - not continuous everywhere, 165
 - with centering, 173
- conditional prob. measure, 14
- confidence set, 74, 76
- conjugate, *see* adjoint
- conservation laws, 52
- continuity
 - of addition
 - for M_1 , 135, 136
 - for M_2 , 159
 - of composition, 165
 - of multidim. reflect.
 - in uniform top., 202
 - on (D, SJ_1) , 203
 - with M_1 tops., 210
 - of the inverse map, 184
 - right, 114
- continuous-mapping approach, 17, 51, 73
- continuous-mapping theorem, 17, 19
- convergence
 - characterization of
 - SM_1 , 128, 138
 - SM_2 , 149, 160
 - WM_1 , 131
 - WM_2 , 155
 - extending to product spaces
 - for SM_1 , 134
 - for SM_2 , 158
 - local uniform, 124
 - of prob. measures, 1
 - of restrictions, 143
 - of sets, 246
 - strengthening the mode
 - for WM_1 , 134
 - for WM_2 , 158
 - to Lévy processes, 45
- convergence preservation
 - WM_2 within bnding fcts., 161
 - with centering
 - inverse map, 188
 - reflection map, 182
 - supremum map, 174, 180
- coordinate mapping, 10
- corrections, v
- counterexample
 - for weak consistency, 94
- counting fcts., 190–194
 - asym. equiv., inv. fcts., 192, 193
- counting functions, 55
 - CLT equivalence, 62
 - with centering, 62
- counting process, 190
- couplings, 27
- covariance function, 34
- Crane, 88
- Csörgő, 25, 27–29
- cumulative process, 37
- cylinder set, 15
- D , the space, 161
 - SM_2 and WM_2 tops., 144
 - characterization of
 - M_1 convergence, 128
 - M_2 convergence, 148
 - regularity properties, 114
- $D([0, \infty), \mathbb{R}^k)$, 142, 216
- D_c , piecewise-const. fcts., 116
- D_0 , subset with $x(0) \geq 0$, 164
- $D_\uparrow$, nondecreasing x in D_0 , 164

- $D_{\uparrow\uparrow}$, increasing x in $D_{\uparrow}$, 164
- D_m , x^i monotone, all i , 164
- D_u , x in D_0 unbded above, 183
- $D_{u\uparrow} \equiv D_u \cap D_{\uparrow}$, 183
- $D_{u\uparrow\uparrow} \equiv D_u \cap D_{\uparrow\uparrow}$, 183
- $D_{u,\epsilon}$, subset of D_u , 184
- D_u^* , subset of D_u , 184
- $D_{u\uparrow}^*$, subset of $D_{u\uparrow}$, 184
- D_s , jumps same sign, 205, 215
- D_1 , jumps in one coord., 214
- D_+ , jumps up, 215
- D_l , piecewise-linear, 208
- $D_{s,l} \equiv D_s \cap D_l$, 208
- D_{lim} , fcts. with left and rt. lims., 243
- $D_{l,r}$, limits either left or rt., 248
- $Disc(x)$, set of disc. pts., 19, 115
- D_Q , subset of D with discs. at rationals, 25
- $d_{J_1}(x_1, x_2)$, J_1 metric, 25
- $d_{M_1}(x_1, x_2)$, M_1 metric, 25, 118
- $d_s(x_1, x_2)$, SM_1 metric on D , 118
- $d_w(x_1, x_2)$, WM_1 dist. on D , 118
- $\hat{d}(A, \Gamma_x)$, order-consist. dist., 119
- $d_p(x_1, x_2)$, product metric, 121
- $d^*(A, A_n)$, graph subsets, 129
- $\hat{d}(A, G_x)$, order-const. dist., 131
- $d_{s,2}(x_1, x_2)$, SM_2 dist., 148
- $d_{w,2}(x_1, x_2)$, WM_2 dist., 148
- Damerdjii, 90
- derivative, v
 - and conv. preservation, 237
 - of inverse map, 267
 - of supremum map, 243
 - of the reflection map, 259
- difference operator
 - fractional, 48
- differences, 47
- Dirac measure, 10
- discontinuity points, 115
- jumps common sign, 136, 159
- distance
 - in-probability, 17
 - order-consistent, 119, 131
- Donsker's theorem
 - rate of convergence in, 25
- double sequences, 41
- Dudley, 6
- Dunford, 16
- Edgeworth expansion, 24
- Egoroff's theorem, 12
- Einmahl, 27
- El-Taha, 52, 54, 71, 72, 164
- equicontinuous, 2
- errors, v
- estimating a steady-state mean, 89
- estimation process, 74
- Ethier, 29, 40
- extending
 - conv. to product spaces
 - for SM_1 , 134
 - for SM_2 , 158
 - graphs for M' tops., 186
- FARIMA, 47
- FCLT
 - for a CTMC, 34
 - for a DTMC, 31
 - for regenerative processes, 38
 - martingale, 40
 - with weak dependence, 30
- feasible regulator
 - definition, 197
- Feller, 22, 24, 41, 44, 45, 140, 230
- Fishman, 75, 91
- fixed-point char. of ref. map, 199
- Fox, 75, 87, 91
- fractional AR integrated MA, FARIMA, 47

- fractional difference operator, 48
- function
 - oscillation, 125
 - piecewise-constant, 115
- functions of sample means, 88
- fundamental matrix
 - for a CTMC, 33
 - of a DTMC, 31
- G_x , thick graph, 117
- gamma process, 100
- generalized cont.-map. thm., 20
- generalized Pollaczek-Khintchine trans-
 - form, 22, 98, 99
- Glazebrook, 87
- Gleser, 87
- Glynn, iv, 27, 36, 38, 53, 71–95
- Gnedenko, 41
- Goldie, 81
- Govindarajulu, 87
- Granger, 48
- graph
 - extended for M' topologies, 186
 - thick, 117
 - thin, 117
- Gut, 71, 72
- Halmos, 12
- Hausdorff metric
 - on compact subsets of $\mathbb{R}_+$, 246
 - on graphs for D , 144
- heavy-traffic limit
 - for nonstationary queues, 262
- heavy-traffic limits
 - for finite-capacity queues, 217
 - for queueing networks, 217
 - for stochastic fluid networks, 217
- Heidelberger, 75, 89
- Helly selection theorem, 140
- Hill estimator, 92
- Horváth, 25, 27
- Hosking, 48
- Hsu, 27
- Iglehart, 90
- in-probability distance, 17
- infinitely divisible distribution, 41
- infinitesimal generator matrix for
 - CTMC, 33
- inheritance of jumps from WM_2
 - convergence, 157
- innovation process, 47
 - heavy-tailed, 50
- instantaneous reflection map, 204
- Internet Supplement, 164, 167, 210, 220
- inverse
 - map, 183–194
 - continuity of, 184
 - conv. pres. with centering, 188
 - derivative of, 267
 - relation for cting. fcts., 55, 190
- J , max-jump fct., 157
- jackknife, 88
- Jacod, 40–42, 45, 46, 142, 178
- Jenkins, 47, 48
- joint conv. of ran. elts.
 - for sup with centering, 178
- Joyeux, 48
- Kella, v, 98, 218, 224
- Kelton, 75, 87, 91
- Kemeny, 31, 33
- Kennedy, 26
- Kiefer-Wolfowitz stochastic approx., 91
- Koenig, 87

- Kogan, 111
- Kolmogorov, 41
- Komlós, 27
- Kurtz, 29, 40
- $L = \lambda W$, 52, 71
- Lévy
 - exponent, 42
 - measure, 42
 - metric, 18, 140
 - process
 - convergence to, 41
 - decomposition, 43
 - model for input, 223
 - reflected, 98
 - strong approx. for, 29
 - without negative jumps, 98
- L_1 topology on D , 210
- $Linc(x)$, left inc. pts. of x , 248
- Laplace exponent, 99
- Lavenberg, 75, 88, 110
- Law, 75, 87, 91
- left limits, 114
- Lemoine, 88
- limiting stationary version
 - criterion for a, 221–223
 - of a reflected process, 218
- Lindvall, 27, 142
- linear
 - fcts. of coord. fcts., 136, 159
 - models, 46
 - process representation, 46, 50
- Lipschitz
 - function, 181
 - mapping theorem, 17, 26
 - property of
 - multidim. ref. map, 202, 203, 215, 217
 - lin. fct. of coord. fcts., 135, 158
- Liptser, 111
- Little's law, 52
- local uniform convergence, 124
- local-maximum function
 - for M_2 top. on D , 160
- Lucantoni, 22, 39
- M/G/1 queue, 99
- M/M/1 queue, 35, 262
- M_1 metric
 - on D , 118
- M'_1 top. on $D([0, \infty), \mathbb{R})$, 186
- M_2 metric
 - on D , 144
- M'_2 top. on $D([0, \infty), \mathbb{R})$, 186
- M_{t_1, t_2} , local max fct., 160
- $m_s(x_1, x_2)$, SM_2 metric on D , 144
- $m_p(x_1, x_2)$, WM_2 product metric on D , 145
- Maigret, 36
- Major, 27
- Mandelbaum, v, 27, 112, 235, 276
- Markov chains, 30
- martingale, 40
 - difference, 40
 - in Lévy pr. decomp., 43
- Massey, v, 71, 72, 112, 235, 276
- matrix
 - norm, 200
 - reflection, 197
- maximum
 - norm on $\mathbb{R}^k$, 114
- McLeish, 92
- measurability
 - of C in (D, M_1) , 164
 - of $Disc(x)$, 19
 - of addition on $D \times D$, 136
 - of subsets of D , 184
 - of the inverse map, 184
- metric

- M_1 , 118
 - SM_1 , 118
 - SM_2 , 145
 - WM_2 , 145
 - Hausdorff, 144
 - Lévy, 140
 - product, 145
 - uniform on D , 114
- Meyn, 36
- Miller, 89
- mixture, 11
- modulus of continuity
 - over a set, 116
- modulus of continuity of BM, 28
- $M_t/M_t/1$ queue, 262
- multidimensional reflection, 233
 - counterexamples for M_1 , 217
 - as function of
 - the ref. matrix Q , 203
 - characterization of
 - by complementarity, 199
 - by fixed-point property, 199
 - continuity of
 - in M_1 , 210
 - as fct of x and Q , 216
 - definition of, 197
 - existence of, 198
 - instantaneous, 204
 - Lipschitz property
 - for $D([0, \infty), \mathbb{R}^k)$, 217
 - for uniform norm, 202
 - for M_1 , 215
 - for SJ_1 , 203
 - one-sided bounds, 202
 - properties of, 202
- multiprogrammed computer system
 - input-output devices, 110
- $N(m, \sigma^2)$, normal variable, 23
- Nadas, 87
- nested family of countable partitions, 7
- Neuts, 39
- Neveu, 15
- non-compact domains, 142
- nonlinear centering, 237
- nonstationary queues, 262
- norm
 - matrix, 200
 - maximum on $\mathbb{R}^k$, 114
- number of visits to a strip, 137
- numerical transform inversion, 22, 48, 100
- order
 - consistent distance, 119, 131
 - on completed graph, 117
 - total, 117
- oscillation function, 125
- parametric representation
 - M_1
 - definitions, 118
 - M_2 , 148
 - limits for, 209
 - reflection of, 204
- partition, 5
 - of a set, 7
- Philipp, 27
- piecewise
 - constant fct., 115
 - linear fct., 208
- planning queueing simulations, 73
- Poisson
 - equation, 31
 - for a CTMC, 34
 - for a DTMC, 32
 - random measure, 43
- Prabhu, 99, 100
- preservation

- of par. rep. by reflection, 205
 - of ptwise. converg., 51
- product
 - measure, 10
 - metric, 145
 - space, 10
- Prohorov, 1–6, 139
- Prohorov metric, 2, 17, 25
- projection map, 26
- Puhalskii, v, 97, 105, 108, 110, 111, 186, 277
- $Q \equiv \lim_{t \downarrow 0} (P(t) - I)$, 33
- $Q \equiv P^t$, reflection matrix, 197
- queueing network, 233
- queueing networks, 217
- Révész, 27–29
- $R \equiv (\psi, \phi)$, mult. ref. map, 197
- $Rinc(x)$, rt. inc. pts. of x , 248
- r_{t_1, t_2} , restriction map, 143
- radius of a set, 8
- random
 - measure, 43
 - sum, 37, 164
 - time change, 72, 164
- range
 - thick, 118
 - thin, 117
- rates of convergence
 - for heavy-traf. limits, 26
 - in CLT, 24
 - in FCLT, 25
- reflected Lévy process, 22, 98
 - steady-state distribution, 98
- reflected process
 - limiting stationary version, 218
 - tightness of marginals, 221
- reflection
 - matrix, 197
 - norm, 200
 - of a parametric rep., 204
- reflection map, 26
 - multidimensional, *see* multidim
 - one-sided, one-dim., 181–183
 - M_2 -cont. fails, 181
 - conv. pres. with ctring., 182
 - derivative of, 259
 - Lipschitz property, 181
- reflexive space, 17
- regenerative
 - cycles, 36
 - process, 36
 - structure, 30
- regularly varying, *see* Appendix A
 - function, 52
- Reiman, 112
- Reinsel, 47, 48
- remainder processes, 37
- renewal process, 37
- renewal-reward processes, 194
- rescaling of mult. ref. map, 202
- Resnick, 92, 93
- restriction of fct., 143
- Robbins, 87
- Robbins-Monro stochastic approx., 91
- Rogers, 100
- Ruppert, 91, 92
- SM_1
 - converg. charact. of
 - main theorem, 128
 - by linear maps, 136
 - visits to strips, 138
 - metric, 118
- SM_2
 - converg. charact. of
 - main theorem, 149
 - by extrema, 160

- by linear maps, 159
 - metric, 145
 - param. rep., 148
 - topology, 144
- Samorodnitsky, 47, 50, 100
- sample mean
 - functions of, 88
 - of IID random variables, 86
 - of IID random vectors, 87
 - with infinite variance, 93
- sample-path method, 22
- Sauer, 75, 88, 110
- Schrage, 75, 87
- Schwartz, 16
- segment
 - product, 117
 - standard, 117
- semimartingales, 46
- separable metric space, 1
- sequential stopping rule
 - absolute-precision, 77
 - relative-precision, 78
- service interruptions, 217
- set
 - feasible regulator, 197
 - of discontinuity points, 115
- Shiryaev, 40–42, 45, 46, 111, 142, 178
- Siegmund, 87
- signed measures, 16
- Simmons, 16
- simulation
 - application of limits to, 73
 - run length, 74
 - sequential stopping rules
 - asymptotic validity, 73
- Skorohod, 6–16, 29, 45, 121, 128, 129, 137, 144, 149, 155, 160, 161, 194
 - embedding theorem, 29
 - representation theorem, 2, 6
- SLLN and FSLLN equivalence, 53
- Smorodinskii, 111
- Snell, 31, 33
- space
 - topological
 - of sets, 246
 - topologically complete, 139
- Srivastava, 87
- Stărică, 92, 93
- stable innovations, 50
- Starr, 87
- stationary
 - process, 218
- steady-state distribution
 - of a reflected Lévy pr., 98
- steady-state mean
 - estimating, 89
- Stidham, 52, 54, 71, 72, 164
- stochastic approximation
 - Kiefer-Wolfowitz, 91
 - Robbins-Monro, 91
- stochastic fluid networks, 217
- Stolyar, 112
- Stout, 27
- Strassen representation theorem, 17
- strengthening mode conv., 134
- strong
 - approximations, 27
 - dependence, 46
- supremum map, 54, 173–180
 - conv. pres., centering, 174
 - criterion for joint conv., 178
 - ctring in other direction, 180
 - derivative of, 243
- switching u and r in par. rep. of
 - inverse fct., 184
- Takács, 98, 100
- Taqqu, 47, 50, 100

- Teugels, 81
- theorem
 - Berry-Esseen, 24
 - Borel-Cantelli, 28
 - continuous-mapping, 17, 19
 - Donsker's, 25
 - Egoroff's, 12
 - gen. cont.-mapping, 20
 - Helly selection, 140
 - Lipschitz-mapping, 17, 26
 - Prohorov metric, 2
 - Skorohod embedding, 29
 - Skorohod representation, 2, 6
 - Strassen representation, 17
 - strong approximation, 27
- thick
 - graph, 117
 - range, 118
- thin
 - graph, 117
 - range, 117
- Thorin, 44
- tightness
 - of a reflected process, 221
- topologically complete, 139
- totally skewed Lévy motion, 43
- transform, 22
- triangular arrays, *see* double sequences
- triple of characteristics for Lévy pr., 42
- truncation function for Lévy pr., 42
- Tusnády, 27
- two-sided regulator, 217
- $u(x_1, x_2, t, \delta)$, unif. dist. fct., 124
- uniform
 - convergence of integrals, 2
 - distance functions, 124
 - metric, 114
 - metric for cdf's, 18, 24
 - norm, 16
- uniformly bounded, 2
- upper semicontinuity
 - of max. abs. jump fct., 157
 - preservation by infimum, 198
- useful functions, 194
- $v(x; A)$, modulus of cont., 116
- $v(x_1, x_2, t, \delta)$, unif. dist. fct., 125
- $v_{t_1, t_2}^{a, b}(x)$, visits to strip $[a, b]$, 137
- Venter, 91, 92
- virtual Mark. arriv. pr., *see* BMAP
- volume of confidence set, 74
- WM_1
 - conv. characteriz. of, 131
 - topology, 118
- WM_2
 - convergence
 - characterizations of, 155
 - inherit jumps from, 157
 - pres. within bnding fcts., 161
 - param. rep., 148
 - topology, 144
- $w_s(x, t, \delta)$, SM_1 oscil. fct., 125
- $w_w(x, t, \delta)$, WM_1 oscil. fct., 125
- $\bar{w}_s(x_1, x_2, t, \delta)$, SM_2 osc. fct., 125
- $\bar{w}_w(x_1, x_2, t, \delta)$, WM_2 osc. fct., 125
- $\bar{w}_s(x, \delta)$, SM_2 osc. fct., 128
- $w_w(x, \delta)$, osc. fct., 131
- Wang, 27
- weak
 - consistency
 - a counterexample for, 94
 - convergence, 1, 16
 - dependence, 30
 - topology, 17
- weak* topology, 17

Welch, 75

Wetherill, 87

Wichura, 6

$Y = \lambda X$, 72

$Z \equiv (I - P + \Pi)^{-1}$, for DTMC, 31

Zhang, 27

Zolotarev, 98